A Finite Population Model of Molecular Evolution: Theory and Computation

Narendra M. Dixit
Department of Chemical Engineering
Indian Institute of Science
Bangalore-560 012, India
Email: narendra@chemeng.iisc.ernet.in

Piyush Srivastava
Computer Science Division
University of California at Berkeley
Berkeley - 94720, CA, USA
Email: piyushsr@eecs.berkeley.edu

Nisheeth K. Vishnoi
Microsoft Research
Bangalore - 560 025, India
Email: nisheeth.vishnoi@gmail.com

Abstract

This paper is concerned with the evolution of haploid organisms that reproduce asexually. In a seminal piece of work, Eigen and coauthors proposed the quasispecies model in an attempt to understand such an evolutionary process. Their work has impacted antiviral treatment and vaccine design strategies. Yet, predictions of the quasispecies model are at best viewed as a guideline, primarily because it assumes an infinite population size, whereas realistic population sizes can be quite small. In this paper we consider a population genetics-based model aimed at understanding the evolution of such organisms with *finite* population sizes and present a rigorous study of the convergence and computational issues that arise therein. Our first result is structural and shows that, at any time during the evolution, as the population size tends to infinity, the distribution of genomes predicted by our model converges to that predicted by the quasispecies model. This justifies the continued use of the quasispecies model to derive guidelines for intervention. While the stationary state in the quasispecies model is readily obtained, due to the explosion of the state space in our model, exact computations are prohibitive. Our second set of results are computational in nature and address this issue. We derive conditions on the parameters of evolution under which our stochastic model mixes rapidly. Further, for a class of widely used fitness landscapes we give a fast deterministic algorithm which computes the stationary distribution of our model. These computational tools are expected to serve as a framework for the modeling of strategies for the deployment of mutagenic drugs.

Topics: Molecular Evolution, Quasispecies Theory.

Contents

1 Introduction

The rapid genomic evolution of viruses such as HIV has made the design of drugs and vaccines with lasting activity one of the most difficult challenges of our time. A novel intervention strategy which potentially outplays viruses in this evolutionary game was suggested by the pioneering work of Eigen and coworkers [Eig71, EMS89]. Eigen and coworkers considered the asexual evolution of a haploid organism and found that when the mutation (or evolutionary) rate was small, the organism survived as a collection of closely related yet distinct genomes together termed the quasispecies. Viral populations in infected individuals are known to exist as such quasispecies [LA10]. Remarkably, this quasispecies model predicted that when the mutation rate increased beyond a critical value, called the error threshold, the collection of genomes in the quasispecies ceased to be closely related; in fact, a completely random collection of genomic sequences was predicted to emerge. This transition with increasing mutation rate thus induced a severe loss of genetic information in the quasispecies and has been referred to as an error catastrophe. The generic antiviral drug ribavirin has been shown to act as a *mutagen*–an agent that induces mutations–against poliovirus and trigger a severe loss of viral infectivity in culture [CCA01]. This strategy of enhancing the viral mutation rate thus appears promising and particularly advantageous because it is unlikely to be susceptible to failure through viral evolution-driven development of drug resistance. Mutagenic drugs that attempt to induce an error catastrophe are thus being explored as a potential antiviral strategy [CCA01, GPLL$^+$05, ADL04], and one such drug for HIV is currently under clinical trials [MHH$^+$11].

The success of mutagenic strategies relies on accurate estimates of the error threshold of the pathogens under consideration. Notwithstanding the tremendous insights into viral evolution the quasispecies model provides, important gaps remain between the quasispecies model and the realistic evolution of viruses and other haploid asexual organisms. First, whereas the model assumes an infinite population size and, hence, adopts a deterministic approach, real populations are often small enough to lend themselves to substantial stochastic effects. For instance, the effective population size of HIV-1 in infected individuals is about $10^3 - 10^6$ [KAB06, BSSD11], which is thought to underlie the strongly stochastic nature of HIV-1 evolution. Second, the model assumes a single-peak fitness landscape, where one genomic sequence is assumed to be the fittest and all other genomes are equally less fit. Realistic fitness landscapes are far more complex [HMC$^+$11]. There have been significant efforts in the last 30 years to close these gaps [Wil05]. While more general landscapes have been successfully considered in the quasispecies case [SH06], a rigorous treatment of the finite population case has remained elusive (see Section 4). Importantly, it still remains to be established whether the insights gained from the quasispecies model, such as the occurrence of an error catastrophe, translate to the more realistic, finite population case.

Here, we consider a finite population model of the asexual evolution of a haploid organism. Following standard population genetics-based descriptions [HC06], the model considers evolution in discrete, non-overlapping generations. Within each generation, genomes undergo reproduction (R), selection (S), and mutation (M), yielding progeny genomes for the next generation. We analyze this RSM model formally and establish the following key results. We show that in the infinite population limit, the expected structure of the quasispecies predicted by the RSM model converges to that of the quasispecies model. Thus, insights from the quasispecies model may be translated to the finite population scenario. Indeed, we show further that the error threshold predicted by the RSM model also converges in the infinite population limit to that of the quasispecies model. Finite population models, such as the RSM model, appropriately tuned to mimic specific details, such as the fitness landscape, of the pathogens under consideration may thus be employed to obtain realistic estimates of the error threshold.

Unlike the quasispecies model, where the quasispecies is identified readily using black-box eigenvector

determination algorithms, identifying the expected quasispecies of the RSM model is computationally prohibitive even for the smallest realistic genome and population sizes. Monte Carlo sampling techniques are therefore often resorted to [Wil05, AB05, BKP$^+$11, GD10, TBVD12]. Here, going beyond the ideas of the quasispecies model, we examine the mixing properties of the RSM model. We establish constraints on the model parameters under which the RSM model exhibits rapid mixing and therefore allows fast estimation of the expected quasispecies structure. Finally, we suggest an algorithm that uses the Markov Chain Monte Carlo paradigm to estimate the error threshold in the RSM model. Our study thus serves as a framework for elucidating quantitative guidelines for the modeling of intervention strategies that employ mutagenic drugs.

The paper is organized as follows. In Section 2 we briefly describe the quasispecies model and the notion of the error threshold. In Section 3 we setup the finite population RSM model, present our main results and outline the techniques employed. In Section 4 we discuss our results in the context of previous studies and highlight open problems arising from work. Formal statements of our results are presented in Section 5. Detailed proofs are contained in the Appendix.

2 Preliminaries and Definitions

2.1 Preliminaries

We consider the evolution of a population of haploid organisms that reproduce asexually. In this evolutionary process the genome of each individual is modeled as a string of nucleotides. During reproduction, the genome is copied with possible mutations, which can be insertions, deletions, or substitutions. In applications, such as the modeling of viral evolution, it is often convenient to neglect insertions, deletions, and substitutions other than transitions (*A* to *G* or *C* to *T*, and *vice versa*). Under this assumption, a genome may, without loss of generality, be represented as a binary string. We thus represent a genome as an L-bit string $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_L)$, where $\sigma_i \in \{0,1\}$.

The *fitness* of a genome is then modeled in terms of its propensity to produce copies of itself. Specifically, the fitness of the genome σ is defined by the number of copies a_σ of itself that it produces in one round of replication (also called one generation). However, during replication, each bit of each of the a_σ offsprings is copied incorrectly with probability μ (called the error or mutation *rate*), thus potentially giving rise to an L-bit string different from σ. The fittest genome, also termed the *master sequence*, is without loss of generality assumed to be $\mathbf{0} = (0, \ldots, 0)$ so that $a_{\mathbf{0}} > a_\sigma$ for all $\sigma \neq \mathbf{0}$.

The primary cause of the complexity and diversity in the evolution of such organisms is the variety of possible fitness landscapes, which *a priori* can be arbitrary functions from $\{0,1\}^L$ to the set of non-negative integers. Several special classes of fitness landscapes have been employed in the literature and we list the important ones below. We will assume that $a_\sigma \geq 1$. The case $a_\sigma = 0$ for some σ's has been used [WK93, TH07, GD10], and will be discussed in Section 4.

1. (*General*) Here the only condition is that $a_\sigma \geq 1$.

2. (*Class Invariant [SH06, TH07, PMnD10, BSSD11]*) In a class invariant landscape a_σ depends only on the Hamming weight of σ.

3. (*Single Peak [Eig71, NS89, PMnD10]*) Here, we have $a_{\mathbf{0}} > 1$ and $a_\sigma = 1$ for all $\sigma \neq \mathbf{0}$.

4. (*Multiplicative [TH07, WH96]*) These are parametrized by $a_1, \ldots, a_L \geq 1$ so that for a given σ, $a_\sigma \stackrel{\text{def}}{=} \prod_{i=1,\sigma_i=0}^{L} a_i$.

Other landscapes such as the simpler additive or linear landscapes and more complex correlated landscapes have also been employed in the literature [BS93, Wie97, vNCM99].

2.2 The Quasispecies Model

Eigen and coworkers [Eig71, EMS89] gave the following differential equations for the time-evolution of the fractional population of the genome σ at time t, denoted by $x_\sigma(t)$:

$$\frac{\mathrm{d}x_\sigma(t)}{\mathrm{d}t} = \sum_\tau x_\tau(t) a_\tau Q_{\tau\sigma} - x_\sigma(t)\bar{A}(t) \text{ for all } \sigma.$$

Here, $Q_{\sigma\tau} \stackrel{\text{def}}{=} \mu^{d_H(\sigma,\tau)}(1-\mu)^{L-d_H(\sigma,\tau)}$ is the probability that σ mutates to τ and $d_H(\sigma,\tau)$ is the Hamming distance between σ and τ. $\bar{A}(t)$ is the average fitness $\sum_\sigma a_\sigma x_\sigma(t)$ at time t. Defining $A_{\sigma\tau} \stackrel{\text{def}}{=} a_\sigma$ when $\sigma = \tau$ and 0 otherwise, they showed that the vector of stationary frequencies, $v_\mu^\sigma \stackrel{\text{def}}{=} \lim_{t\to\infty} x_\sigma(t)$, is the dominant right eigenvector of the *value matrix QA* at mutation rate μ such that $\|\mathbf{v}_\mu\|_1 \stackrel{\text{def}}{=} \sum_\sigma v_\mu^\sigma = 1$.[1] The collection of genomes determined by this dominant eigenvector, which marks the culmination of the evolutionary process, is called the *quasispecies*. It is important to note that the vector $\mathbf{v}_\mu$ is independent of the starting population distribution.

We will mostly be concerned with the discrete time version of the quasispecies model. In the discrete time case, $t = \{0,1,\ldots\}$, denoting the fraction of genomes of type σ at time t by m_t^σ, Eigen's equations can be written as:

$$m_{t+1}^\sigma \stackrel{\text{def}}{=} \frac{\sum_\tau m_t^\tau a_\tau Q_{\tau\sigma}}{\sum_\tau m_t^\tau a_\tau}. \tag{1}$$

In vector notation, given the fractional population $\mathbf{m}_t$ at time t, the fractional population $\mathbf{m}_{t+1}$ at time $t+1$ is given by $\mathbf{m}_{t+1} = \mathbf{r}(\mathbf{m}_t)$, where the σ co-ordinate r^σ of the vector valued function $\mathbf{r}$ is defined as

$$r^\sigma(\mathbf{x}) \stackrel{\text{def}}{=} \frac{\sum_\tau a_\tau Q_{\tau\sigma} x_\tau}{\sum_\tau a_\tau x_\tau} = \frac{(QA\mathbf{x})_\sigma}{||A\mathbf{x}||_1} \text{ and thus, } \mathbf{r}(\mathbf{x}) = \frac{QA\mathbf{x}}{||A\mathbf{x}||_1}. \tag{2}$$

Again, $\mathbf{m}_t$ can be shown to converge to $\mathbf{v}_\mu$ irrespective of the starting population distribution as t goes to infinity. However, at any finite t, $\mathbf{m}_t$ depends on the initial state $\mathbf{m}_0$.

2.3 The Error Threshold

With the single peak landscape, Eigen *et al.* observed empirically that there is a critical value $\mu_c \leq 0.5$ for the mutation rate μ such that for $\mu \ll \mu_c$, the quasispecies is dominated by the master sequence, i.e. $v_\mu^{\mathbf{0}} \geq v_\mu^\sigma \ \forall\sigma$, whereas when $0.5 \geq \mu > \mu_c$ the quasispecies is dispersed approximately uniformly. The critical mutation rate μ_c is called the *error threshold* because the uniform dispersal for $\mu > \mu_c$ implies a severe loss of representation in the quasispecies of the genetic information encoded by the master sequence. Evidently, this dispersal also decreases the mean fitness, $\overline{A} \stackrel{\text{def}}{=} \sum_\sigma a_\sigma v_\mu^\sigma$.

We note here that despite the notion of the error threshold being widely recognized, no consensus exists on its definition; see Wilke [Wil05]. Since, in most cases, $\mathbf{v}_\mu$ will never become exactly the uniform distribution on $\{0,1\}^L$, it is clear that the goal is to find the smallest μ such that $\mathbf{v}_\mu$ is *close* to the uniform distribution on all genomes, i.e., the vector with every coordinate equal to 2^{-L}, which we denote by $\mathbf{U}$. To

[1] Throughout this paper, we will be dealing with vectors over $\{0,1\}^L$. Vectors will be typeset in boldface. The components of a vector $\mathbf{x}$ will be denoted either by x^σ or by x_σ for $\sigma \in \{0,1\}^L$, based on convenience of notation in the context of use.

define the error threshold we also need a function that measures closeness: e.g. $\|\mathbf{v}_\mu - \mathbf{U}\|_1$, $\|\mathbf{v}_\mu - \mathbf{U}\|_\infty$ or the difference in Shannon entropies of $\mathbf{v}_\mu$ and $\mathbf{U}$, namely $|\mathbf{H}(\mathbf{v}_\mu) - L|$. Hence, for a given distance function d which measures closeness of $\mathbf{v}_\mu$ and $\mathbf{U}$, and a bound on closeness $\varepsilon > 0$, we define

$$\mu_c^d(\varepsilon) \stackrel{\text{def}}{=} \min\{\mu \in (0,1) : d(\mathbf{v}_\mu, \mathbf{U}) \leq \varepsilon\}.$$

First note that at $\mu = 1/2$, the steady state vector $\mathbf{v}_\mu$ is *exactly* $\mathbf{U}$. Hence, $\mu_c^d(\varepsilon) \leq 1/2$ for all $d, \varepsilon > 0$. Second, note that changing the distance function d will change the error threshold quantitatively. Eigen and coworkers presented a heuristic argument that the error threshold should scale as $1/L$ for the single-peak model without any rigorous proofs of its existence and without mentioning any closeness function.

3 A Finite Population Model and Our Main Results

In this section, we describe at an informal level the salient features of our work, which comprises a finite population model to capture molecular evolution, and our theoretical and computational results associated with it. We give a high-level technical overview of the methods used to prove our results in Section 3.3, while precise definitions and formal statements of our results appear in Section 5. Proofs have been moved to the Appendix due to considerations of space.

3.1 A Finite Population (RSM) Model

We consider the following stochastic discrete time finite population model of evolution which we call the RSM model. The parameters are the same as in the quasispecies model: the genome length L, the sequence space $\{0,1\}^L$, a per bit mutation rate μ and the fitness landscape $\{a_\sigma\}_{\sigma \in \{0,1\}^L}$ with all $a_\sigma \geq 1$ and integers. At any time t, let N_t^σ be the number of genomes (a random variable) of type σ, and fix the total population $\sum_\sigma N_t^\sigma$ to be N. This fixing of the population size to N at each step is the key distinction from the quasispecies model and is a new parameter. In each time step, the ensuing evolution is then described in terms of the following three steps.

1. (**Reproduction**) First, in the reproduction step, each genome σ produces a_σ copies of itself, giving rise to an intermediate population $I_t \stackrel{\text{def}}{=} \sum_{\sigma \in \{0,1\}^L} I_t^\sigma$, where $I_t^\sigma \stackrel{\text{def}}{=} a_\sigma N_t^\sigma$.

2. (**Selection**) Second, in the selection step, N genomes are chosen at random without replacement from this intermediate population of size I_t, resulting in the selection of S_t^σ genomes of type σ where $S_t^\sigma \in \{0,1,\ldots,I_t^\sigma\}$ and $\sum_{\sigma \in \{0,1\}^L} S_t^\sigma = N \leq I_t$.

3. (**Mutation**) Third, in the mutation step, each selected genome is mutated with probability μ per bit, giving rise to the next generation of N_{t+1}^σ genomes of type σ, such that $\sum_{\sigma \in \{0,1\}^L} N_{t+1}^\sigma = N$.

The starting state is denoted by $\mathbf{N}_0$ which is typically ggiven by $N_0^{\mathbf{0}} = N$ and $N_0^\sigma = 0$ for all $\sigma \neq \mathbf{0}$, but we will often not use this assumption. This RSM model is best viewed as a Markov chain where the state space is the set of functions $f : \{0,1\}^{2^L} \mapsto \{0,1,\ldots,N\}$ such that $\sum_\sigma f(\sigma) = N$. Thus, the number of states of this Markov chain is $\binom{N+2^L-1}{N}$ which is roughly N^{2^L}. It can be shown (see Fact 5.1) that for any $0 < \mu < 1$ the transition matrix of this Markov chain has a unique stationary vector, denoted by π. π is indexed by all f satisfying the property above and $\sum_f \pi(f) = 1$, i.e., π is a probability distribution over the state space of the RSM Markov chain.

Let $\mathbf{D}_t \stackrel{\text{def}}{=} (N_t^\sigma/N)_{\sigma\in\{0,1\}^L}$ denote the random vector which captures the fractional population at time t. It can also be shown that $\lim_{t\to\infty}\mathbb{E}[\mathbf{D}_t|\mathbf{D}_0]$ exists and is independent of $\mathbf{D}_0$. We denote this limit as $\mathbb{E}[\mathbf{D}_\infty]$ and it can be computed from the stationary vector π as follows:

$$\mathbb{E}[D_\infty^\sigma] = \sum_{i=0}^{N}\sum_{f:f(\sigma)=i}\frac{i}{N}\cdot\pi(f).$$

Finally, we will subscript $\mathbf{D}_t$ with parameters such as μ, N when we want to highlight dependence on them, e.g. $\mathbf{D}_{t,\mu,N}$. The key questions of interest, especially given the fact that computing π would be prohibitive even for small values of L and N, are:

1. For a fixed t, what does $\mathbb{E}[\mathbf{D}_t|\mathbf{D}_0]$ converge to as N increases?
2. Is there a notion of error threshold in the RSM model?
3. How to obtain an estimate of $\mathbb{E}[\mathbf{D}_\infty]$ efficiently?

We present theoretical results that address all of the above questions. Note that if the answer to the first question is that $\mathbb{E}[\mathbf{D}_t|\mathbf{D}_0]$ converges to the prediction of the quasispecies model $\mathbf{m}_t$ with the same starting states ($\mathbf{m}_0 = \mathbf{D}_0$), then it is important as we can leverage the significant understanding obtained from the study of the quasispecies model while incorporating stochastic effects with finite populations. For the second question, we need to first define a notion of the error threshold in the RSM model. We do so formally, given a distance function d.

Definition 3.1. *Let* $\varepsilon \geq 0$.

$$\mu_c^d(\varepsilon, N) \stackrel{\text{def}}{=} \min\left\{\mu \in (0,1) : d(\mathbb{E}\left[\mathbf{D}_{\infty,\mu,N}\right], \mathbf{U}) \leq \varepsilon\right\},$$

where $\mathbf{U}$ *is the uniform distribution over all genomes of length* L.

What we want to understand is $\lim_{N\to\infty}\mu_c^d(\varepsilon,N)$. Again, if we can show that the answer here is $\mu_c(\varepsilon)$, then we can translate the insights from the quasispecies model to the RSM model.

For the third question, first note that if we want to estimate the error threshold, we need to be able to compute $\mathbb{E}[\mathbf{D}_\infty]$. Secondly, we consider the standard notion of efficiency: the algorithm to estimate $\mathbb{E}[\mathbf{D}_\infty]$ should be polynomial in the input size. As we noted, the state space of the RSM Markov chain is prohibitively large and computing the stationary state is prohibitive. We employ the Markov Chain Monte Carlo method and run the RSM process for some time τ such that it is guaranteed that distribution from which $\mathbf{D}_\tau$ is drawn comes *statistically close* to the distribution from which $\mathbf{D}_\infty$ is drawn, irrespective of $\mathbf{D}_0$. Simulating each step of the random walk can be done efficiently. Hence, we are led to the question of bounding the *mixing time* of the RSM Markov chain: the smallest time the finite time distribution needs to come close to the steady state distribution for all starting configurations.

The issue of how the input is presented is also important and we briefly discuss it here. In one model, one can be given all a_σ which would require bit length about $\sum_\sigma \log a_\sigma$ and can, in principle, be as large or even larger than 2^L. Often, this is not the case and either the values a_σ are given by a simple equation, or only some fixed number, say $k \ll 2^L$ of the values a_σ are strictly bigger than 1. In the latter case the input has bit length roughly $O(k\log\max_\sigma a_\sigma + \log L + \log 1/\mu)$. Another case for the input is when a_σ are class invariant. In this case the input is of length roughly $O(L\log\max_\sigma a_\sigma + \log 1/\mu)$. We now proceed to summarize our results.

3.2 Our Results

We now give informal statements of our main results, before describing the mathematical techniques employed in the proofs of these results. The formal statements of the theorems described here appear in Section 5, after a discussion in Section 4, while the formal proofs are deferred to the Appendix.

3.2.1 Convergence of the Quasispecies and the RSM Model

Theorem 3.2 (Convergence of the RSM and Quasispecies Models)**.** *Fix a fitness landscape A with positive entries and a mutation transition matrix Q. Consider the RSM process started with the initial state* $\mathbf{D}_0$ *and consider the evolution of the quasispecies model started with the initial state* $\mathbf{m}_0 = \mathbf{D}_0$*. Then for any fixed time t,*

$$\lim_{N\to\infty} \mathbb{E}\left[\mathbf{D}_t|\mathbf{D}_0\right] = \mathbf{m}_t, \tag{3}$$

where $\mathbf{m}_t$ *is the fractional population vector at time t starting from* $\mathbf{m}_0$ *predicted by the quasispecies model.*

The theorem shows that in the infinite population limit, the stochastic fluctuations of the RSM process disappear, and the model converges to the quasispecies model. Informally, the main technical difficulty in proving the above theorem is to establish a result of the form $\lim_{N\to\infty}\mathbb{E}[\mathbf{D}_t|\mathbf{D}_{t-1}] = \lim_{N\to\infty}\mathbb{E}\left[\mathbf{D}_t|\mathbf{D}_0\right]$ with probability 1, which would establish convergence to the quasispecies model. The full proof is deferred to the Appendix. This convergence result allows us to show that for any distance function d, a finitary version of the error threshold, $\mu_c^d(\varepsilon,N)$ as defined above, converges to the error threshold $\mu_c^d(\varepsilon)$ of the quasispecies model, as the population size goes to infinity. These two results provide validation for the finite population RSM model by establishing that in the infinite population limit, the predictions from the RSM model converge to those of the quasispecies model. We now move on to problems concerning the mixing time and other computational issues of the RSM model.

3.2.2 Computational Results in the RSM Model

As noted before, a primary computational question in both the quasispecies model and the RSM model is the determination of the quasispecies, or the expected population profile at stationarity in the RSM model, which can then be used to estimate error thresholds (see Section 3.3 for an overview and Sections 5.3.2 and 5.3.3 for details). For the quasispecies model, a satisfactory solution to this problem is obtained via the observation that the quasispecies is the leading right eigenvector of the QA matrix. The QA matrix is of dimension $2^L \times 2^L$, and the above observation can thus be used to obtain efficient algorithms using black-box eigenvector finding algorithms for moderate values of L. In the case of class-invariant fitness landscapes, it is known [SS82] that one only needs to find the leading eigenvector of an $(L+1)\times(L+1)$ matrix.

However, similar approaches are not as effective for the RSM model. In this case, the stationary distribution is the leading eigenvector of the transition matrix $\mathscr{M}$ of the RSM process which is of dimension roughly N^{2^L}. Using ideas similar to those referred to above, one can reduce the running time for computing the stationary distribution to $N^{O(L^2)}$.

Theorem 3.3 (Computation of Steady State in the Class Invariant Case)**.** *For any class invariant fitness landscape A, there is an algorithm running in time* $T = O(N^{O(L^2)})$ *which computes the steady state of the RSM process with population size N and the genome length L.*

However, as mentioned before, in many applications, as in the case of HIV, for instance, where $N \sim 10^3 - 10^6$ and $L \sim 10^4$, the problem is still computationally prohibitive. In these cases, one typically resorts to Monte Carlo simulations of the RSM process for estimating the population profile at stationarity [TBVD12], and thus we are led to considering the mixing time of the RSM process. The following theorem derives conditions on the parameters of evolution under which the RSM model mixes rapidly.

Theorem 3.4 (Mixing Time of the RSM Process)**.** *Given a fitness landscape A, mutation rate μ, the RSM process exhibits fast mixing if* $(1-2\mu)\frac{\max_\tau a_\tau}{\min_\tau a_\tau} L + \frac{1}{N} < 1$.

Having stated our results, we now highlight the techniques employed in the proofs.

3.3 Overview of Our Technical Contributions

As before, we will denote by $\mathbf{D}_t$ the random variable of fractional populations after t steps of the RSM process, and by $\mathbf{S}_t$ the random variable of the populations of genomes after the replication and selection steps of the $(t+1)$-the step of the RSM process.

Our convergence result (Theorem 3.2). The starting point of the proof of our convergence result is to observe that $\mathbb{E}[\mathbf{D}_{t+1}|\mathbf{D}_t]$ has the same functional form $\mathbf{r}$ (as a function of $\mathbf{D}_t$) as the evolution equation of the discrete time quasispecies model, with $\mathbf{r}$ as defined in Equation (2): $\mathbb{E}[\mathbf{D}_{t+1}|\mathbf{D}_t] = \mathbf{r}(\mathbf{D}_t)$. Our high level approach is to first show that $\mathbf{D}_{t+1}$ is actually *concentrated* around $\mathbb{E}[\mathbf{D}_{t+1}|\mathbf{D}_t]$. Using the Lipschitz continuity of the evolution function $\mathbf{r}$, we can then chain these concentration results inductively to show that the evolution of $\mathbf{D}_t$ is tightly concentrated around the evolution of the discrete time quasispecies model, which allows us to show that $\mathbb{E}[\mathbf{D}_t]$ converges to the quasispecies as $N \to \infty$. To illustrate the ideas involved, we consider the case $L = 1$. Here the two genomes are $\{0,1\}$. After the replication phase in the t-th step, there are $a_0 D_t^0 N$ copies of 0. For the i-th copy, let R_i denote the indicator variable for this copy being selected in the selection phase, so that $S_t^0 = \sum_{i=1}^{a_0 D_t^0 N} R_i$. Since the R_i's are not independent, we cannot directly apply a Chernoff bound. However, since they are negatively correlated, one expects concentration to hold, and this can indeed be shown using the so-called method of bounded differences. The same reasoning works for S_t^1, and thus we get that given $\mathbf{D}_t$, the intermediate population $\mathbf{S_t}$ after the replication and selection steps is concentrated around its expectation with high probability. We now look at the mutation step. Let M_i be the indicator variable for the ith genome being $\mathbf{0}$ after the mutation step. We then have $N D_{t+1}^0 = \sum_{i=1}^N M_i$. Since the M_i's are independent random variables, it can be shown using a Chernoff bound that given $\mathbf{S}_t$, D_{t+1}^0 is concentrated around $\mathbb{E}\left[D_{t+1}^0|\mathbf{S}_t\right] = 1/N(\mu S_t^0 + (1-\mu)S_t^1)$. The two steps can then can be combined to show that given $\mathbf{D}_t$, D_{t+1}^0 is concentrated around $\mathbb{E}\left[\mathbb{E}\left[D_{t+1}^0|\mathbf{S_t}\right]|\mathbf{D}_t\right] = 1/N\mathbb{E}\left[\mu S_t^0 + (1-\mu)S_t^1|\mathbf{D}_t\right] = \frac{(1-\mu)a_0 D_t^0 + \mu D_t^1}{a_0 D_t^0 + D_t^1}$. The same reasoning works for D_{t+1}^1.

With some more work, this argument can be generalized to work for arbitrary L. The concentration guarantee we obtain is of the following form: there are quantities ε_t and p_t which are both $o_N(1)$ such that given $\mathbf{D}_t$, $|\mathbf{D}_{t+1} - \mathbb{E}[\mathbf{D}_{t+1}|\mathbf{D}_t]| \le \varepsilon_t$ with probability at least $1 - p_t$. In the next step, we chain these stepwise bounds inductively in order to remove the conditioning and show that for all $t \le t_0$, $\mathbf{D}_t$ is concentrated around $\mathbf{m}_t$. An important component of the induction is the observation that $\mathbf{r}$ is Lipschitz continuous, which allows us to control the propagation of the errors ε_t' in each step. By the induction hypothesis, we have that $|\mathbf{D}_t - \mathbf{m}_t| \le \varepsilon_t'$ with probability at least $1 - p_t'$, where ε_t' and p_t' are both $o_N(1)$. Assuming the Lipschitz constant of $\mathbf{r}$ to be K, this implies that $\mathbb{E}[\mathbf{D}_{t+1}|\mathbf{D}_t] = \mathbf{r}(\mathbf{D}_t)$ is within distance $K\varepsilon_t'$ of $\mathbf{m}_{t+1} = \mathbf{r}(\mathbf{m}_t)$ with probability at least $1 - p_t'$. Applying the convergence result from the first step, we then have that with probability at least $1 - p_{t+1}' = 1 - p_t' - p_t$, $|\mathbf{D}_{t+1} - \mathbf{m}_{t+1}| \le \varepsilon_{t+1}' = K\varepsilon_t' + \varepsilon_t$ of $\mathbf{m}_{t+1}$. The quantities p_t', ε_t' for

$t \leq t_0$ can be chosen to be $o_N(1)$, which is sufficient to show the required convergence. The details appear in Appendix B.2. We now give an overview of the proofs of our computational results.

Computing the stationary distribution in the class invariant case (Theorem 3.3). Recall that the state space of the RSM Markov chain is roughly N^{2^L}. However, if the fitness function is class invariant, we can show that the number of *distinct* coordinates in the state space is about N^L. To see this, first we define an equivalence relation on the states of the RSM Markov chain. We say that f, g, which are functions from $\{0,1\}^L$ to $\{0,1,\ldots,N\}$ satisfying $\sum_\sigma f(\sigma) = N$ and $\sum_\sigma g(\sigma) = N$, are *equivalent* (denoted $f \equiv g$) if they have the same statistics for every Hamming class, i.e., for every $0 \leq i \leq N$,

$$\sum_{\left\{\sigma \in \{0,1\}^L | w_H(\sigma) = i\right\}} f(\sigma) = \sum_{\left\{\sigma \in \{0,1\}^L | w_H(\sigma) = i\right\}} g(\sigma).$$

Thus, the state space of the RSM Markov chain gets partitioned into about $(N+1)^{L+1}$ different classes. Then, due to the fact that the fitness function is class invariant, it can be shown that the transition probability of f to any other equivalence class is the same as that of g to the same class. Hence, one only needs to compute the transition probability from one equivalence class to another. This probability is a large binomial sum and one has to be careful in its computation and keep track of the number of bits required to represent each entry of this Markov chain over the equivalence classes. Once we have the transition matrix of this Markov chain, one can compute its largest eigenvector which corresponds to the stationary state. We show that, if one does this carefully, one can compute the eigenvector in time roughly $N^{O(L^2)}$. The details appear in Section B.5.

Algorithm to compute the error threshold. Once we have the ability to either compute the stationary state of the RSM process or derive independent samples from its stationary state (which allows us to estimate the relative frequencies of the genomes at stationarity with a good precision by taking an average of the sampled states), the algorithm to estimate the error threshold is simple. The idea is to start with a small value ($\ll 1/L$) of μ, and to estimate/compute the stationary distribution of the RSM process for the current value of μ. The algorithm then checks if the estimate of the stationary distribution is close to the uniform distribution on the genomes in the measure of closeness of one's choice. If so, it stops and outputs the current value of μ as an estimate of the error threshold. Else, it increases μ by a very small amount and repeats the above steps. In case direct computation of the stationary state of the RSM process is computationally prohibitive, independent samples from the stationary distribution of the RSM process are derived by simulating the RSM process up to its mixing time. The number of samples required can be estimated from a simple application of Chernoff bound on the random variable corresponding to the stationary state distribution of the RSM process. Hence, to establish bonds on the running time of the error threshold estimating algorithm, it is important to be able to bound the time it takes for the RSM process so that $\mathbf{D}_t$ comes close to the stationary state, $\mathbf{D}_\infty$. Our next result is towards this.

Mixing time result (Theorem 3.4). Since the stationary distribution of the RSM chain is not very well understood, it is not clear how to apply conductance-based geometric tools or the canonical paths method (see, for example, [JS88]) in order to prove the mixing time result. We are thus led to more combinatorial coupling based methods. Here one starts with an integer valued metric d on the state space of the Markov chain, and then one runs two copies X_t and Y_t of the chain. To show fast mixing, it is then sufficient to prove that X_t and Y_t can be coupled so that $\mathbb{E}\left[d(X_{t+1}, Y_{t+1}) | (X_t, Y_t)\right] \leq \alpha < 1$. In general, defining a coupling can

be tricky because one needs to carefully argue that the marginals of the coupling agree with the original Markov chain.

We define the coupling in two phases: the first phase includes the replication and selection steps and the second phase includes only the mutation step. We begin with the easier mutation step. Let I_t and J_t denote the state of the two RSM chains after the replication and selection steps. For most natural choices of the distance metric d, it is possible to couple the mutation step using the standard coupling for the random walk on the hypercube so that $\mathbb{E}[d(X_{t+1},Y_{t+1})|(I_t,J_t)] \leq (1-2\mu)d(I_t,J_t)$. The challenge however lies in controlling $\mathbb{E}[d(I_t,J_t)|(X_t,Y_t)]$ while coupling the replication and selection steps, since because of the global nature of the replication and selection steps, $\mathbb{E}[d(I_t,J_t)|(X_t,Y_t)]$ can become quite large. We control this increase by a careful choice of the metric d, and by appealing to the path coupling methods of Bubley and Dyer [BD97]. The path coupling theorem says that for integer valued d, it is sufficient to ensure $\mathbb{E}[d(X_{t+1},Y_{t+1})|(X_t,Y_t)] \leq \alpha \leq 1$ *only* for states X_t and Y_t satisfying $d(X_t,Y_t) = 1$ in order to establish fast mixing. Our coupling is then defined as follows. Fix a permutation of the N genomes in the chain X_t: $d(X_t,Y_t)$ is then the minimum over all possible permutations of the N genomes in Y_t of the sum of the Hamming distances between the genomes at the same positions in the two permutations. The main technical step is to show that for this d, the replication and selection steps can be coupled in such a way that starting from X_t and Y_t satisfying $d(X_t,Y_t) = 1$, $\mathbb{E}[d(I_t,J_t)|(X_t,Y_t)]$ after these steps is at most $\frac{N}{N-1}\frac{\max_\sigma a_\sigma}{\min_\tau a_\tau}L$. It is in this step that we use the form of the distance metric d crucially; the details of the coupling are somewhat technical and involve arguing carefully that the coupling is valid, and are given in Appendix B.4. We then combine this with the coupling for the mutation step described above to show contraction in the expected distance under the condition $(1-2\mu)\frac{N}{N-1}\frac{\max_\sigma a_\sigma}{\min_\tau a_\tau}L < 1$.

4 Discussion and Future Perspectives

4.1 Previous Work

The notion of the quasispecies and the existence of an error threshold were recognized first by Eigen and his coworkers in the 1970s and 1980s [Eig71, EMS89]. Translation of these ideas into intervention strategies requires overcoming two key limitations of the quasispecies model. First, the model assumes an infinite population size, whereas realistic population sizes can be quite small. With HIV, for instance, the effective population size is estimated to be $\sim 10^3 - 10^6$ [KAB06, BSSD11]. Second, the theory assumes a single-peak fitness landscape, whereas realistic landscapes can be far more complex [BCP$^+$04, HMC$^+$11]. Efforts over the last several decades have attempted to overcome these limitations of the quasispecies model [NS89, BS93, WH96, Wie97, AF98, SH06, TH07, PMnD10] (see Wilke [Wil05] for a recent review). The finite population case, however, has remained difficult to solve in full generality. Most studies resort to simulations or use approximate or heuristic approaches to describe the finite population case, and we discuss some of these here.

Nowak and Schuster [NS89] used a birth-death process to model the underlying evolution in finite populations and using simulations predicted that the error threshold scales as $1/\sqrt{N}$. Their model, however, does not converge to the quasispecies model as N goes to infinity. Alves and Fontanari [AF98] present a model which employs a two-stage sampling with replacement in the selection process: first sampling uniformly from the population, and then sampling from the obtained sample with biases proportional to the fitness. They note, however, that sampling with replacement destroys the negative correlation between the selection of two individuals of the same species induced by the finite population constraint when selection is implemented using sampling *without* replacement. They find that the error threshold scales as $1/N$.

Further, they only analyze a heuristic deterministic approximation to their model, and do not consider rigorously the question of convergence of their original model to the quasispecies model. The closest to our convergence result is that by van Nimwegen *et al*. [vNCM99] who show convergence to a deterministic Eigen-like dynamics but again employ sampling with replacement and use special cases of additive fitness landscapes. With finite populations, there can be a significant statistical difference between sampling with and without replacement, the latter (which we employ) being more realistic. It is well known that as $N \to \infty$ the difference between sampling with and without replacement shrinks, but then as we prove, so does the difference between our population genetics model and the quasispecies model. Their convergence proof has a similar structure as ours but we are able to use Chernoff bound type inequalities which are much stronger than the second moment inequalities used by them. Consequently, our convergence results are quantitatively stronger. We additionally prove convergence of the error threshold and fast mixing, questions not considered by [vNCM99]. More recently, Musso [Mus11] presented the transition matrix for sampling with replacement in the case $L = 1$ and also claimed convergence to the quasispecies model in the deterministic limit. No attempt, however, is made in [Mus11] to make this latter claim rigorous. Another class of studies relies on approximations and heuristics inspired from physics, and in particular statistical mechanics, to render the finite population case mathematically tractable (e.g., [BK98, SRA08, PMnD10]).

While previous studies have focused extensively on the fractional distribution of genomes at stationarity, little is known of the time to reach the stationary state. Campos and Fontanari [CF99] show that in the limit of infinitely large genome lengths ($L \to \infty$) and population sizes ($N \to \infty$) and with the single peak fitness landscape, the timescale associated with the decline of the master sequence is $1/\ln(qa)$ where q is the probability that a genome is replicated without error, and a is the relative fitness of the master sequence. Further, they show that with finite populations, this timescale is proportional to $\sqrt{N}$. The mixing time when L and N are both finite and when the fitness landscape is more general than the single peak remains unknown. The latter mixing time has practical significance in the modeling of the action of mutagenic drugs, as it respresents the duration of therapy required to ensure completion of the transition to the error catastrophe. Our study presents conditions when the mixing is rapid and hence the transition to error catastrophe occurs quickly. Further, for computational studies that attempt to realize this transition *in silico*, our study presents an algorithm that allows efficient Monte Carlo sampling-based estimation of the error threshold.

4.2 Applications of the RSM Model

The motivation behind the RSM model and the algorithms discussed here is to get a basic framework for understanding the evolution of viruses of current interest such as HIV. Making concrete predictions relevant to the clinical setting requires super-imposing the specifics of the viruses of concern on the present framework. This often involves subtle modifications of the RSM process along with validation against data. For example, in related recent work, two of the authors and their co-workers applied the RSM model to mimic the within-host genomic evolution of HIV-1 [TBVD12]. It has been shown before [BSSD11] that these simulations quantitatively capture data of the evolution of viral genomic diversity in patients over extended durations (~ 10 years) following infection and the approach is extended in [TBVD12] to estimate the error threshold of HIV-1. We envision that similar adaptation of our model will prove useful in elucidating the evolution and treatment guidelines for other asexual haploid organisms of interest.

4.3 Critique of the RSM Model

We note that our structural and computational results are independent of the nature of the fitness landscape so long as there are no lethal mutations ($a_\sigma \neq 0$ for any σ). Our model, however, does not consider lethal

mutations. While letting some a_σ be 0 does not affect the quasispecies model due to the constant rescaling involved, it introduces an absorbing state in the RSM Markov chain, thus making it non-ergodic, and causing the population to eventually decrease to zero. While Wilke and others [Wil05, WK93, TH07] have commented on the role of lethal mutations in extreme cases, establishing their full implications lies beyond the scope of the present paper. Although lethal mutations do occur, it turns out that in many important scenarios such as the evolution of HIV-1, a Hamming class invariant landscape without lethal mutations appears to capture key features of the underlying fitness interactions [BCP$^+$04], rendering our RSM framework applicable.

Finally, we note that our assumption of a fixed population size, N, is consistent with the widely accepted population genetics-based models of evolution, where a constant effective population size is employed to quantify the strength of stochastic effects [HC06]. Note that allowing N to vary with time (generations), does not increase the complexity in our model. The distinction between an infinite population model and a finite population model arises from the culling of the population in the latter model in order to maintain a finite population size. A fixed N or varying N will only result in different extents of culling in different generations, but will not change the overall structure of the model. The advantage in keeping N constant for our present study is that it allows easier examination of the convergence to the quasispecies model.

4.4 Open Problems

Our study of the quasispecies and RSM models has revealed several interesting and important problems. We list the main ones here.

Structure of the Quasispecies. Perhaps the most attractive feature of finite population models as opposed to the quasispecies model is that they can be used to study the effect of random genetic drift on inter-patient variations. Inter-patient variations in disease progression and response to treatments are known to be significant with HIV infection [NBS$^+$98, GKB$^+$05]. The collection of viral particles in an infected individual may be thought of as one realization of the random viral evolutionary process, and $\lim_{t\to\infty} \mathbf{Var}\,[D_t^\sigma]$ then provides an estimate of inter-patient variations in viral evolution due to the effect of the finite population size. Thus, in addition to the structure of the quasispecies in the finite population case, defined by the expected frequencies $\lim_{t\to\infty} \mathbb{E}\,[\mathbf{D}_t]$ when $N < \infty$, the variance of the frequencies $\lim_{t\to\infty} \mathbf{Var}\,[D_t^\sigma]$ as a function of the population size N is also an important quantity to be studied.

Error Threshold. In the quasispecies model with the single-peak fitness landscape, μ_c has been found, without a rigorous proof, to be $O(1/L)$, so that an error catastrophe occurs for $\mu \ll 0.5$ (e.g., see [EMS89]). Further, the transition is sharp, so that a small increase in μ from below to above μ_c induces a dramatic change in the quasispecies structure. With other fitness landscapes, such as the multiplicative landscape, however, the quasispecies approaches the uniform distribution gradually as μ approaches 0.5 [WH96]. Further, lethal mutations, where $a_\sigma = 0$ for some σ's, appear to show the existence of an error threshold only if multiple mutations in a single replication are allowed [WK93, TH07, Wil05]. Thus, the conditions under which a sharp transition leading to an error catastrophe at $\mu_c \ll 0.5$ would occur remain to be established. Second, the dependence of μ_c on N remains to be identified. While some simulations suggest a $1/\sqrt{N}$ dependence [NS89, BS93], others find the dependence to go as $1/N$ [AF98]. As pointed out before, knowledge of μ_c for finite N is important in the modeling of antiviral strategies based on mutagenic drugs.

Mixing Time. The main outstanding question here is to get a tight bound on the mixing time of the RSM Markov chain for a full range of evolutionary parameters. We notice that our result shows a good mixing time bound only under certain conditions on the parameters. Though we conjecture that the chain is rapidly mixing for other values of the parameters too, we believe that novel methods would be needed to extend our results in this direction. Apart from being useful in determining the time required for simulations to produce samples from the stationary distribution, the mixing time bounds also have biological significance. For example, when modeling the effect of a mutagenic drug under the RSM model, the convergence rate would models the minimum required duration of treatment before the error catastrophe occurs.

5 Formal Statements of Main Results

5.1 Preliminaries and Definitions

In this section we present rigorous statements of our results. Several definitions may be found repeated here in the interest of the readability of this section. We recall that genomes of length L are denoted by L-bit 0-1 strings. We will denote the Hamming distance between genomes σ and τ by $d_H(\sigma,\tau)$, and the Hamming weight of a genome σ by $w_H(\sigma)$. A *population* is defined as a multiset of genomes of the same length. While discussing the RSM model, we will fix the size of the population to be N.

The Markov Chain for the RSM Model. We will denote the evolution of the RSM process using a time-indexed sequence of vector valued random variables $(\mathbf{N}_t)_{t=0}^{\infty}$. The entries of $\mathbf{N}_t$ are indexed by genomes σ, and the entry N_t^σ denotes the number of genomes of type σ at time t. At every time t, $\sum_{\sigma\in\{0,1\}^L} N_t^\sigma = N$. The random variables $D_t^\sigma \stackrel{\text{def}}{=} N_t^\sigma/N$ denote the *fractional population* of the genome σ at time t.

Reproduction Step and the Fitness Landscape In the *reproduction* step, each genome σ produces a_σ copies of itself, so that the number of genomes of type σ after this step is $I_t^\sigma \stackrel{\text{def}}{=} a_\sigma N_t^\sigma$, and the total number of genomes is $I_t \stackrel{\text{def}}{=} \sum_\sigma a_\sigma N_t^\sigma$. The matrix A defined by $A_{\sigma\sigma} \stackrel{\text{def}}{=} a_\sigma$ and $A_{\sigma\tau} \stackrel{\text{def}}{=} 0$ for $\sigma \neq \tau$ is called the *fitness landscape*. The fitness landscape is said to be *class-invariant* if a_σ depends only on the Hamming weight of σ. By a slight abuse of notation, we will denote by a_i the fitness of all genomes with Hamming weight i in the class invariant case.

Selection and Mutation Steps and the Mutation Rate In the *selection* step, N genomes are sampled without replacement from the genomes obtained after the reproduction step. In the *mutation* step, each bit of each of the N genomes obtained after the selection step is flipped with a probability μ, called the *mutation rate*. The *mutation transition matrix* Q defined by $Q_{\sigma\tau} \stackrel{\text{def}}{=} \mu^{d_H(\sigma,\tau)}(1-\mu)^{L-d_H(\sigma,\tau)}$ gives the probability that a genome of type σ mutates to one of type τ in the mutation step.

The RSM process as described above is a Markov chain on the state space of functions $f : \{0,1\}^L \longrightarrow \mathbb{N}$, satisfying $\sum_{\sigma\in\{0,1\}^L} f(\sigma) = N$. The transition matrix $\mathscr{M}$ of this chain is described in Section A.

Fact 5.1. *When the mutation rate $\mu \in (0,1)$ and $a_\sigma > 0$ for all σ, the Markov chain $\mathscr{M}$ corresponding to the RSM process is ergodic, and hence has a unique stationary distribution.*

This is a simple consequence of the fact that μ and A are positive. See Section B.1 for a proof.

Important Statistics of the RSM Process and the Projected RSM Process. The transition matrix $\mathscr{M}$ of the RSM Markov chain is of dimension $\binom{N+2^L-1}{N} \times \binom{N+2^L-1}{N}$. When the fitness landscape A is class invariant, one can get a projected Markov chain with a significantly smaller state space which can still be used to compute the average fitness and the average population of each fitness class at stationarity. Consider equivalence classes indexed by functions $h : [0,L] \longrightarrow N$ with $\sum_{i=0}^{L} h(i) = N$, such that a function f in the state space of $\mathscr{M}$ is in the equivalence class $[h]$ if and only if for every $i \in [0,L]$, $\sum_{\{\sigma \in \{0,1\}^L | w_H(\sigma)=i\}} f(\sigma) = h(i)$. We then have the following lemma the proof of which is in Section B.1.

Lemma 5.2. *Let f,g belong to the same equivalence class h as defined above, and let $[h']$ be another equivalence class. We then have $\mathscr{M}(f,[h']) = \mathscr{M}(g,[h'])$.*

Thus, we can consider the projected Markov chain $\mathscr{M}_w$ with state space

$$\Omega_w = \left\{ [h] \middle| \sum_{i=0}^{L} h(i) = N \right\}.$$

Notice that $|\Omega_w| = \binom{N+L}{L}$. Also, if π_w is the stationary distribution of $\mathscr{M}_w$, then by the projection property

$$\pi_w([h]) = \sum_{f \in [h]} \pi(f).$$

This property implies that the expected populations for every Hamming class of genomes and, hence, the expected average fitness at stationarity, are the same for $\mathscr{M}_w$ and $\mathscr{M}$.

Mixing Time. We will denote by π the stationary distribution of the RSM process, and let $\mathbf{N}_\infty$ be a random variable distributed according to π. We know that the distributions of the random variables $\mathbf{N}_t$ converge in total variation distance (and hence in distribution) to π, due to the ergodicity of the RSM process. We fix our notation for mixing times in this section.

Definition 5.3. *The total variation distance between two probability distributions $\mathscr{D}_1$ and $\mathscr{D}_2$ on the sample space Ω is defined by $\|\mathscr{D}_1 - \mathscr{D}_2\|_{TV} = \max_{A \subseteq \Omega} |\mathscr{D}_1(A) - \mathscr{D}_2(A)|$.*

Definition 5.4. *Fix a Markov chain $\mathscr{N}$ on a state space S. We define*

$$d(t) \stackrel{\text{def}}{=} \max_{\alpha \in S} \|\mathscr{N}^t(\alpha,\cdot) - \pi\|_{TV}.$$

For $0 \leq \varepsilon \leq 1/2$, the mixing time of $\mathscr{N}$ is defined by

$$\tau_{\text{mix}}(\varepsilon) \stackrel{\text{def}}{=} \min\{t : d(t) \leq \varepsilon\}.$$

Notice that by the projection property, the mixing time of the projected RSM chain $\mathscr{M}_w$ is at most the mixing time of the original RSM chain $\mathscr{M}$.

Error Thresholds. In the following definition, we specifically emphasize the dependence of the random variables $\mathbf{N}_t, \mathbf{D}_t$ on μ, N by denoting them as $\mathbf{N}_{t,\mu,N}$ and $\mathbf{D}_{t,\mu,N}$. We denote by $\mathbf{D}_{\infty,\mu,N}$ a version of $\mathbf{D}_{t,\mu,N}$ distributed according to the stationary distribution of the RSM process, and by $\mathbf{U}$ the uniform distribution over genomes. Given a distance function d, one can define the error threshold with respect to d as follows.

Definition 5.5 (Error Threshold for the RSM Model)**.** *Let* $\varepsilon \geq 0$.

$$\mu_c^d(\varepsilon, N) \stackrel{\text{def}}{=} \min\left\{\mu \in (0,1) : d(\mathbb{E}\left[\mathbf{D}_{\infty,\mu,N}\right], \mathbf{U}) \leq \varepsilon\right\},$$

where $\mathbf{U}$ *is the uniform distribution over all genomes of length* L.

Of particular interest is the function d^h : for any two distributions $\mathscr{D}_1$ and $\mathscr{D}_2$ over genomes, $d^h(\mathscr{D}_1, \mathscr{D}_2)$ denotes $|\sum_\sigma w(\sigma)(\mathscr{D}_1(\sigma) - \mathscr{D}_2(\sigma))|$. The corresponding μ will carry the superscript h.

5.2 Convergence to the Quasispecies Model

Our first main result is that the RSM model converges to the quasispecies model.

Theorem 5.6. *Fix a fitness landscape* A *with positive entries and a mutation transition matrix* Q*. Consider the RSM started with the initial state* $\mathbf{D}_0$ *and consider the evolution of the quasispecies model started with the initial state* $\mathbf{m}_0 = \mathbf{D}_0$*. Then for any fixed time* t_0*,*

$$\lim_{N\to\infty} \mathbb{E}\left[\mathbf{D}_{t_0} | \mathbf{D}_0\right] = \mathbf{m}_{t_0}, \tag{4}$$

where $\mathbf{m}_{t_0}$ *is the state of evolution of the quasispecies model at time* t_0 *starting from* $\mathbf{m}_0$.

The proof of the above theorem is relegated to Section B.2. As a corollary to the theorem above, one can show that there is convergence of a finitary version of the error threshold $\mu_c^h(\varepsilon, N)$ to the error threshold $\mu_c^h(\varepsilon)$ for the quasispecies model, as the population size goes to infinity. Formally, we have the following:

Corollary 5.7. *Fix a mutation rate* $\mu \leq 1/2$ *and an error parameter* ε*. For every* $\delta > 0$*, there exists a time* $t_0 > 0$ *such that for* $t > t_0$*, one can find an* $N_{\delta,t}$ *such that for* $N > N_{\delta,t}$*,*

$$\begin{aligned} d^h\left(\mathbb{E}\left[\mathbf{D}_{t,\mu,N}\right], \mathbf{U}\right) &\geq \varepsilon - \delta, \quad \text{when } \mu < \mu_c^h(\varepsilon), \text{ and} \\ d^h\left(\mathbb{E}\left[\mathbf{D}_{t,\mu,N}\right], \mathbf{U}\right) &\leq \varepsilon + \delta \qquad \text{when } \mu = \mu_c^h(\varepsilon), \end{aligned}$$

Here we use the subscripts μ *and* N *to emphasize the dependence of the distribution of* $\mathbf{D}_t$ *on* μ *and* N.

Although we will prove our results for the error threshold in terms of the average Hamming distance, it is easy to translate our results to other common dispersal measures as described in Section B.1.1. The proof of the above Corollary follows easily from Theorem 5.6 and is given in Section B.3. We note here that extending the above corollary to get convergence of finite population error thresholds depends upon proving a strengthened version of our convergence result (Theorem 5.6), which we leave as an open problem. In fact, on the basis of simulation results, we conjecture that for fixed ε, $\mu_c^h(\varepsilon, N)$ monotonically increases to $\mu_c^h(\varepsilon)$.

5.3 Computational Results

5.3.1 Mixing Time Bounds on the RSM Process

We give a coupling argument in Section B.4 which allows us to prove the following theorem.

Theorem 5.8. *Fix* $0 < \mu \leq 1/2$*, and a fitness landscape* A*. Let*

$$K(A,\mu) \stackrel{\text{def}}{=} (1-2\mu)\frac{N}{N-1}\frac{\max_\sigma a_\sigma}{\min_\tau a_\tau} L.$$

When $K(A,\mu) < 1$*, we have* $\tau_{\text{mix}}(\varepsilon) = O\left(\frac{\log(NL/\varepsilon)}{\log(1/K)}\right)$.

5.3.2 Computing the Stationary Distribution in the Class Invariant Case

Theorem 5.9. *For every A which is class invariant, and μ which can be represented using b bits, there is an algorithm running in time T given by*

$$T \stackrel{\text{def}}{=} \tilde{O}\left(bL^3 + \left(\frac{(N+L)^L}{L!}\right)^3 (NbL + L^2 + N\max a_\sigma) + NbL\left(\frac{(N+L)^L}{L!}\right)^2 \left(e(1+\frac{N}{L(L+1)})\right)^{L(L+1)}\right)$$

which computes π_w for the Markov chain $\mathscr{M}_w$ described above. For fixed L, $T = O(N^{O(L^2)})$.

The proof of the above theorem appears in Section B.5, and is based on the projected RSM process discussed in Section 5.1. The above theorem immediately gives a grid-search based algorithm that given a grid resolution δ and $\varepsilon > 0$, outputs a approximation μ_0 to the error threshold in time $T \cdot 1/2\delta$ such that $\mu_0 \geq \mu_c^h(\varepsilon, N)$ and $d^h(\mathbf{D}_{\infty,\mu_0-\delta}, \mathbf{U}) > \varepsilon$. We now consider Markov Chain Monte Carlo based grid-search methods.

5.3.3 Markov Chain Monte Carlo Methods

The general strategy for Monte Carlo based grid search methods for determining error thresholds is described in the algorithm ERRORTHRESHOLD in Figure 1 in the Appendix. We will denote the mixing time $\tau_{\text{mix}}(\varepsilon)$ for parameters L, N, A and μ as $\tau(L, N, A, \mu, \varepsilon)$. We consider the projected chain $\mathscr{M}_w$ described above which contains enough information to compute the average Hamming weight, and whose state can be maintained as a tuple in $\{0, 1, \ldots, N\}^{L+1}$.

Theorem 5.10. *Let A be class invariant, and consider the error threshold $\mu_c^h(\varepsilon, N)$. Suppose the algorithm* ERRORTHRESHOLD *is run with input grid resolution δ, accuracy parameter δ_1, and error probability δ_2. Let T be the maximum over k of the quantity $\tau(L, N, A, k\delta, \delta_1/(2L))$ where $k \leq 1/(2\delta)$ is a positive integer. The algorithm* ERRORTHRESHOLD *runs in time $T \cdot s \cdot \tilde{O}(\lceil 1/(2\delta)\rceil NL \max_\sigma a_\sigma)$, where*

$$s = \left\lceil \frac{8L^4}{\delta_1^2} \log\left(2\lceil 1/(2\delta)\rceil (L+1)/\delta_2\right) \right\rceil,$$

and with probability at least $1 - \delta_2$, produces an output μ_0 satisfying $\mu_0 \geq \mu_c^h(\varepsilon + \delta_1/2)$ and $d^h(\mathbf{D}_{\infty,\mu_0-\delta}, \mathbf{U}) \geq \varepsilon - \delta_1/2$.

The proof of the above theorem appears in Section B.6, where we also point out some technical subtleties about the definition of error thresholds.

Acknowledgments. We thank Rajesh Balagam for helping us with several simulations based on which this theoretical study was initiated. This work was initiated while Piyush Srivastava was an intern at Microsoft Reseach, Bangalore.

References

[AB05] Christian L. Althaus and Sebastian Bonhoeffer. Stochastic Interplay between Mutation and Recombination during the Acquisition of Drug Resistance Mutations in Human Immunodeficiency Virus Type 1. *Journal of Virology*, 79(21):13572–13578, 2005.

[ADL04] Jon P. Anderson, Richard Daifuku, and Lawrence A. Loeb. Viral error catastrophe by mutagenic nucleosides. *Annual Review of Microbiology*, 58(1):183–205, 2004.

[AF98] D. Alves and J. F. Fontanari. Error threshold in finite populations. *Physical Review E*, 57(6):7008 – 7013, June 1998.

[BCP+04] Sebastian Bonhoeffer, Colombe Chappey, Neil T. Parkin, Jeanette M. Whitcomb, and Christos J. Petropoulos. Evidence for positive epistasis in HIV-1. *Science*, 306(5701):1547–1550, 2004.

[BD97] R. Bubley and M. E. Dyer. Path coupling: a technique for proving rapid mixing in markov chains. In *Proceedings of the 38th IEEE Symposium on the Foundations of Computer Science(FOCS)*, pages 223 – 231, 1997.

[BK98] D. Bonnaz and A. J. Koch. Stochastic model of evolving populations. *Journal of Physics A: Mathematical and General*, 31(2):417, 1998.

[BKP+11] Rebecca Batorsky, Mary F. Kearney, Sarah E. Palmer, Frank Maldarelli, Igor M. Rouzine, and John M. Coffin. Estimate of effective recombination rate and average selection coefficient for HIV in chronic infection. *Proceedings of the National Academy of Sciences*, 108(14):5661–5666, 2011.

[BS93] Sebastian Bonhoeffer and Peter F. Stadler. Error thresholds on correlated fitness landscapes. *Journal of Theoretical Biology*, 164(3):359 – 372, 1993.

[BSSD11] Rajesh Balagam, Vasantika Singh, Aparna Raju Sagi, and Narendra M. Dixit. Taking multiple infections of cells and recombination into account leads to small within-host effective-population-size estimates of HIV-1. *PLoS ONE*, 6(1):e14531, 01 2011.

[CCA01] Shane Crotty, Craig E. Cameron, and Raul Andino. RNA virus error catastrophe: Direct molecular test by using ribavirin. *Proceedings of the National Academy of Sciences*, 98(12):6895–6900, 2001.

[CF99] PRA Campos and JF Fontanari. Finite-size scaling of the error threshold transition in finite populations. *J. Phys A*, 32:L1–L7, 1999.

[DP09] Devdatt P. Dubhashi and Alessandro Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, 2009.

[Eig71] M. Eigen. Selforganization of matter and the evolution of biological macromolecules. *Die Naturwissenschaften*, 58:456–523, 1971.

[EMS89] M. Eigen, J. McCaskill, and P. Schuster. The molecular quasi-species. *Adv. Chem. Phys.*, 75:149–263, 1989.

[GD10] Saikrishna Gadhamsetty and Narendra M. Dixit. Estimating frequencies of minority nevirapine-resistant strains in chronically HIV-1-infected individuals naive to nevirapine by using stochastic simulations and a mathematical model. *J. Virol.*, 84(19):10230–10240, 2010.

[GKB+05] Enrique Gonzalez, Hemant Kulkarni, Hector Bolivar, Andrea Mangano, Racquel Sanchez, Gabriel Catano, Robert J. Nibbs, Barry I. Freedman, Marlon P. Quinones, Michael J. Bamshad, Krishna K. Murthy, Brad H. Rovin, William Bradley, Robert A. Clark, Stephanie A. Anderson, Robert J. O'Connell, Brian K. Agan, Seema S. Ahuja, Rosa Bologna, Luisa Sen, Matthew J. Dolan, and Sunil K. Ahuja. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science*, 307(5714):1434–1440, 2005.

[GPLL+05] Ana Grande-Pérez, Ester Lázaro, Pedro Lowenstein, Esteban Domingo, and Susanna C. Manrubia. Suppression of viral infectivity through lethal defection. *Proceedings of the National Academy of Sciences of the United States of America*, 102(12):4448–4452, 2005.

[HC06] Daniel L. Hartl and Andrew G. Clark. *Principles of Population Genetics, Fourth Edition*. Sinauer Associates, Inc., 4th edition, December 2006.

[HMC+11] Trevor Hinkley, Joao Martins, Colombe Chappey, Mojgan Haddad, Eric Stawiski, Jeannette M Whitcomb, Christos J Petropoulos, and Sebastian Bonhoeffer. A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. *Nature Genetics*, 43:487–489, 2011.

[JS88] Mark Jerrum and Alistair Sinclair. Conductance and the rapid mixing property for markov chains: the approximation of the permanent resolved. In *Proceedings of the 20th Annual ACM Symposium on Theory of Computing (STOC 1988)*, pages 235–243, 1988.

[KAB06] Roger D. Kouyos, Christian L. Althaus, and Sebastian Bonhoeffer. Stochastic or deterministic: what is the effective population size of HIV-1? *Trends in Microbiology*, 14(12):507 – 511, 2006.

[LA10] Adam S. Lauring and Raul Andino. Quasispecies theory and the behavior of RNA viruses. *PLoS Pathog*, 6(7):e1001005, 07 2010.

[MHH+11] James I. Mullins, Laura Heath, James P. Hughes, Jessica Kicha, Sheila Styrchak, Kim G. Wong, Ushnal Rao, Alexis Hansen, Kevin S. Harris, Jean-Pierre Laurent, Deyu Li, Jeffrey H. Simpson, John M. Essigmann, Lawrence A. Loeb, and Jeffrey Parkins. Mutation of HIV-1 genomes in a clinical population treated with the mutagenic nucleoside kp1461. *PLoS ONE*, 6(1):e15135, 01 2011.

[Mus11] Fabio Musso. A stochastic version of Eigen's model. *Bulletin of Mathematical Biology*, 73:151 – 180, 2011.

[NBS+98] Monique Nijhuis, Charles A. B. Boucher, Pauline Schipper, Thomas Leitner, Rob Schuurman, and Jan Albert. Stochastic processes strongly influence HIV-1 evolution during suboptimal protease-inhibitor therapy. *Proceedings of the National Academy of Sciences*, 95(24):14441–14446, 1998.

[NS89] M. Nowak and P. Schuster. Error thresholds of replication in finite populations-mutation frequencies and the onset of Muller's ratchet. *J. Theor Biol*, 137:375–395, 1989.

[PMnD10] Jeong-Man Park, Enrique Muñoz, and Michael W. Deem. Quasispecies theory for finite populations. *Phys. Rev. E*, 81(1):011902, Jan 2010.

[SH06] David B. Saakian and Chin-Kun Hu. Exact solutions of the Eigen model with general fitness functions and degradation rates. *PNAS*, 103(13):4935 – 4939, 2006.

[SRA08] David B. Saakian, Olga Rozanova, and Andrei Akmetzhanov. Dynamics of the Eigen and the Crow-Kimura models for molecular evolution. *Phys. Rev. E*, 78(4):041908, Oct 2008.

[SS82] J. Swetina and P. Schuster. Self-replication with errors: a model for polynucleotide replication. *Biophys. Chem.*, 16:329–345, 1982.

[TBVD12] Kushal Tripathi, Rajesh Balagam, Nisheeth K. Vishnoi, and Narendra M. Dixit. Stochastic simulations suggest that HIV-1 survives close to its error threshold. *Submitted*, 2012.

[TH07] Nobuto Takeuchi and Paulien Hogeweg. Error-threshold exists in fitness landscapes with lethal mutants. *BMC Evolutionary Biology*, 7(1):15, 2007. A response to Claus Wilke: Quasispecies theory in the context of population genetics, BMC Evol Biol 2005, 5:44.

[vNCM99] Eric van Nimwegen, Japen P. Crutchfield, and Melanie Mitchell. Statistical dynamics of the royal road genetic algorithm. *Theoretical Computer Science*, 229:41 – 102, 1999.

[WH96] G. Woodcock and PG Higgs. Population evolution on a multiplicative single-peak fitness landscape. *J. Theor. Biol.*, 179:61–73, 1996.

[Wie97] Thomas Wiehe. Model dependency of error thresholds: the role of fitness functions and contrasts between the finite and infinite sites models. *Genetics Research*, 69(02):127–136, 1997.

[Wil05] Claus Wilke. Quasispecies theory in the context of population genetics. *BMC Evolutionary Biology*, 5(1):44, 2005.

[WK93] G. P. Wagner and P. Krall. What is the difference between models of error thresholds and Muller's ratchet? *J. Math. Biol.*, 32:33–44, 1993.

A Starting State and Transition Matrix of the RSM Markov Chain

As stated before, the RSM Markov chain starts with the "fittest" possible population with all the weight concentrated on the master sequence, so that $N_0^M = N$ and $N_0^\sigma = 0$ for all $\sigma \neq M$. We now proceed to set up some notation for writing out the transition matrix $\mathscr{M}$.

Definition A.1. ***(Multivariate Geometric distribution).*** *Let $g(\sigma)$ denote the number of genomes of type σ in an urn. Consider the process of choosing, without replacement, N genomes from this urn. Then $P^{hyp}(g \to f; N)$ denotes the probability of obtaining $f(\sigma)$ genomes of type σ for each σ. We have,*

$$P^{hyp}(g \to f; N) \stackrel{\text{def}}{=} \frac{\prod_{\sigma \in \{0,1\}^L} \binom{g(i)}{f(i)}}{\binom{\langle g, \mathbf{1} \rangle}{N}} \tag{5}$$

Definition A.2. ***(Multivariate Binomial Distribution).*** *Let $f(\sigma)$ denote the number of genomes of type σ. Consider a stochastic process in which each genome of type σ independently mutates into a genome τ (possibly equal to σ) with probability $Q(\sigma, \tau)$. We denote by $P^{bin}(f \to D; Q)$ the probability that $D(\sigma, \tau)$ genomes of type σ mutate to type τ under this process. We have*

$$P^{bin}(f \to D; Q) \stackrel{\text{def}}{=} \prod_{\sigma \in \{0,1\}^L} \binom{f(\sigma)}{\left\{ D(\sigma, \tau) | \tau \in \{0,1\}^L \right\}} \prod_{\tau \in \{0,1\}^L} Q(\sigma, \tau)^{D(\sigma, \tau)}$$

We can now write the entries of $\mathscr{M}$. For $f, g : \{0,1\}^L \longrightarrow N$ satisfying $\langle f, \mathbf{1} \rangle = \langle g, \mathbf{1} \rangle = N$, we denote by $\mathscr{M}(f, g)$ the conditional probability of obtaining g starting from f in one step of the RSM process. Given a function $f\{0,1\}^L \longrightarrow N$, we denote by Af the function such that $Af(\sigma) = a_\sigma f(\sigma)$. Then, we have

$$\mathscr{M}(f, g) = \sum_{h : \langle h, \mathbf{1} \rangle = N} P^{\text{hyp}}(Af \to h; N) \sum_{D : \mathbf{1}D = g; D\mathbf{1} = h} P^{\text{bin}}(h \to D; Q),$$

where Q and A are as defined above.

B Proofs Omitted from Section 5

B.1 Proofs Omitted from Section 5.1

Proof Sketch of Fact 5.1. When $\mu \in (0,1)$ and $a_\sigma > 0$ for all σ, it can be verified easily that this chain is irreducible and also has a non-zero self-loop probability at every point in the state space. Thus, the chain is *ergodic* and hence by the Fundamental theorem of Markov chains, has a unique stationary distribution to which it converges as $t \to \infty$. □

We now give a proof of Lemma 5.2.

Proof of Lemma 5.2. We will show that under class invariance, we can project the RSM Markov chain so that its state space consists of equivalence classes indexed by functions $h : [0,L] \longrightarrow N$ with $\sum_{i=0}^{L} h(i) = N$, such that a function f in the state space of $\mathscr{M}$ is in the equivalence class $[h]$ if and only if for every $i \in [0,L]$,

$$\sum_{\left\{\sigma \in \{0,1\}^L | w_H(\sigma) = i\right\}} f(\sigma) = h(i).$$

We will find it convenient to consider the reproduction and selection phases separately from the mutation phase, show that the projection described above can be done for both of them, and then combine the two results using the following general fact about projected Markov chains, the proof of which we include for completeness.

Fact B.1. *Let P and R be the transition kernels of two Markov chains on the same state space Ω, and let S denote the composition PR of the two chains. Suppose that there is a partition of Ω into equivalence classes Ω', such that for any $f \equiv f'$, and any equivalence class $[g]$, we have*

$$P(f,[g]) = P(f',[g]) \text{ and } R(f,[g]) = R(f',[g]).$$

Then, we also have $S(f,[g]) = S(f'([g])$, for all f, f' and g as described above.

Proof. The proof is by direct computation. We have,

$$\begin{aligned}
S(f,[g]) &= \sum_{q' \in \Omega} P(f,q')R(q',[g]) \\
&= \sum_{[q] \in \Omega'} \sum_{q' \in [q]} P(f,q')R(q',[g]) \\
&= \sum_{[q] \in \Omega'} R(q,[g]) \sum_{q' \in [q]} P(f,q') \\
&= \sum_{[q] \in \Omega} R(q,[g])P(f,[q]) \qquad (6) \\
&= \sum_{[q] \in \Omega} R(q,[g])P(f',[q]) \qquad (7)
\end{aligned}$$

Just as in the derivation of equation (6) above, we get $S(f',[g]) = \sum_{[q] \in \Omega'} R(q,[g])P(f',[q])$, and hence, by equation (7), we have $S(f',[g]) = S(f,[g])$, as claimed. □

In order to use the last fact, we now decompose the matrix of the RSM process into the following two Markov chains on Ω:

1. **The Reproduce-Select Chain.** We denote the transition matrix of this chain as P, such that $P(f,g)$ is the probability of obtaining the state g starting from state f after the reproduction and selection phases. Notice that
$$P(f,g) = P^{\text{hyp}}(Af \to g; N).$$
Assume that A is class invariant and let $A(i)$ denote the reproduction rate for a genome of Hamming weight i. For an equivalence class $[h]$ of Ω as defined above, we consider the probability $P(f,[h])$, with $f \in [h']$. By the definition of $P^{\text{hyp}}(\to;)$, this is the probability of drawing $h(i)$ genomes of Hamming weight i for $0 \le i \le L$, when N genomes are drawn without replacement from a bag containing $A(i)\sum_{\sigma:w_H(\sigma)=i} f(\sigma) = A(i)h'(i)$ genomes of weight i. By definition, this probability depends only on h and the equivalence class h' of f, and hence $P(f,[h]) = P(g,[h])$ when A is class invariant and $f \equiv g$.

2. **The Mutation Chain.** We will directly write down the entries $R(f,[h'])$ for the probability of obtaining a state in the equivalence class $[h']$ starting from a state f in the equivalence class $[h]$. We will show now that we can write $R(f,[h'])$ in terms only of h and h', and hence $R(f,[h']) = R(g,[h'])$ for $f \equiv g$. Denote by $\mathscr{Q}_{ij}$ the probability that a string σ of Hamming weight i transforms into some string of Hamming weight j in the mutation step, and notice that this probability is well defined because of the definition of the mutation transition matrix Q. Since $f \in [h]$, there are $h(i)$ strings of Hamming weight i initially, for $0 \le i \le L$. Denote by d_{ij} the number of strings of weight i which transform into strings of weight j in the mutation step. Then, we have
$$R(f,[h']) = \sum_{\substack{d:\sum_j d_{ij}=h(i) \\ \sum_i d_{ij}=h'(j)}} \prod_{0\le i\le L} \binom{h(i)}{\{d_{ij} | 0 \le j \le L\}} \prod_{0\le j\le L} \mathscr{Q}_{ij}^{d_{ij}}. \tag{8}$$
Since $R(f,[h'])$ depends only upon h and h', we get that $R(f,[h']) = R(g,[h'])$ for $f \equiv g$.

Combining the above two discussions and using Fact B.1, we see that when A is class invariant, the transition matrix $\mathscr{M}$ of the RSM process satisfies $\mathscr{M}(f,[h']) = \mathscr{M}(g,[h'])$ whenever $f \equiv g$. This completes the proof of Lemma 5.2. □

B.1.1 Relationships between Error Thresholds

We first define error thresholds according to various dispersal measures.

Definition B.2 (Error Thresholds)**.** *Let $\varepsilon \ge 0$, and $\mathbf{U}$ be the uniform distribution over the set of genomes.*

1. $\mu_c^{\text{ex},1}(\varepsilon,N) \stackrel{\text{def}}{=} \min\left\{\mu \in (0,1) : \left\|\mathbb{E}\left[\mathbf{D}_{\infty,\mu}\right] - \mathbf{U}\right\|_1 \le \varepsilon\right\}$.
2. $\mu_c^{\text{h}}(\varepsilon,N) \stackrel{\text{def}}{=} \min\left\{\mu \in (0,1) : \left|\sum_\sigma w_H(\sigma)\,\mathbb{E}\left[D^\sigma_{\infty,\mu}\right] - 2^{-L}\sum_\sigma w_H(\sigma)\right| \le \varepsilon\right\}$.
3. $\mu_c^{\text{sh}}(\varepsilon,N) \stackrel{\text{def}}{=} \min\left\{\mu \in (0,1) : \left|\mathbf{H}(\mathbb{E}\left[\mathbf{D}_{\infty,\mu}\right]) - \mathbf{H}(\mathbf{U})\right| \le \varepsilon\right\}$. *Here $\mathbf{H}$ denotes the Shannon entropy, using the base e.*

Definition B.3 (Error Threshold for the Quasispecies Model)**.** *Let $\mu \in (0,1)$ and let $\mathbf{v}_\mu$ denote the the stationary expected fraction vector with ℓ_1 norm 1. The error thresholds are defined as follows.*

1. $\mu_c^{\text{ex},1}(\varepsilon) \stackrel{\text{def}}{=} \min\{\mu \in (0,1) : \|\mathbf{v}_\mu - \mathbf{U}\|_1 \le \varepsilon\}$.

2. $\mu_c^{\mathrm{h}}(\varepsilon) \stackrel{\text{def}}{=} \min\{\mu \in (0,1) : |\sum_\sigma v_\mu(\sigma) w_H(\sigma) - 2^{-L}\sum_\sigma w_H(\sigma)| \leq \varepsilon\}$.

3. $\mu_c^{\mathrm{sh}}(\varepsilon) \stackrel{\text{def}}{=} \min\{\mu \in (0,1) : |\mathbf{H}(\mathbf{v}_\mu) - \mathbf{H}(\mathbf{U})| \leq \varepsilon\}$, *where* $\mathbf{H}$ *denotes the Shannon entropy, using the base e.*

Our results are mostly stated in terms of the error threshold μ_c^h. However, we now describe how the different definitions above are related to each other. The following inequalities relate the different distance measures that we have considered. The first of these follows from the definition of the ℓ_1 norm, while the second is the well known Pinsker's inequality relating the ℓ_1 norm to the entropy.

$$\left|\mathbb{E}\left[\sum_\sigma w_H(\sigma) D^\sigma_{\infty,\mu}\right] - \mathbb{E}[\sigma \leftarrow U]\sum_\sigma \frac{w_H(\sigma)}{|\Omega|}\right| \leq L\left|\left|\mathbb{E}\left[\mathbf{D}_{\infty,\mu}\right] - \mathbf{U}\right|\right|_1 \tag{9}$$

$$\left|\left|\mathbb{E}\left[\mathbf{D}_{\infty,\mu}\right] - \mathbf{U}\right|\right|_1 \leq \sqrt{2\left|\mathbf{H}(\mathbb{E}\left[D_{\infty,\mu}\right]) - \mathbf{H}(\mathbf{U})\right|} \tag{10}$$

This gives us the following relationship between the error-thresholds:

$$\begin{aligned} \mu_c^h(\varepsilon,N) &\leq \mu_c^{\mathrm{ex},1}(\varepsilon/L,N) \\ \mu_c^{\mathrm{ex},1}(\varepsilon,N) &\leq \mu_c^{\mathrm{sh}}(\varepsilon^2/2,N) \end{aligned} \tag{11}$$

Using the fact that the distributions involved are defined over a state space of size 2^L, we can show the following weak converse to inequality (10):

$$\left|\mathbf{H}(\mathbb{E}\left[\mathbf{D}_{\infty,\mu}\right]) - \mathbf{H}(\mathbf{U})\right| \leq 2^L\left|\left|\mathbb{E}\left[\mathbf{D}_{\infty,\mu}\right] - \mathbf{U}\right|\right|_1^2.$$

This gives us a further relationship between the error thresholds:

$$\mu_c^{\mathrm{ex},1}(\varepsilon,N) \geq \mu_c^{\mathrm{sh}}(2^L\varepsilon^2,N)$$

However, we notice that one cannot in general close the loop in inequalities (9) and (10) (and hence in inequalities (11)) above. To see this, consider for example the following two distributions P and Q for $L > 1$.

1. P: puts total weight $1-\varepsilon$ on weight 1 strings and weight ε on the string $\mathbf{0}$, so that the average Hamming weight is $1-\varepsilon$.

2. Q: puts total weight $(1-\varepsilon)/L$ on weight L strings, and weight $1-(1-\varepsilon)/L$ the string $\mathbf{0}$, so that the average Hamming weight is still $1-\varepsilon$.

The average Hamming weight in both cases is $1-\varepsilon$, so that in that metric, the distance between P and Q is zero. However, the total variation distance between P and Q is at least $1-\varepsilon$.

B.2 Proof of Theorem 5.6

In the rest of this section, we will use the following concentration inequalities about the multivariate hypergeometric distribution:

Fact B.4. *Consider the hypergeometric distribution $P^{hyp}(g \to f;N)$ defined in equation (5) above. Let D^σ be the random variable denoting the fraction of genomes of type σ which are drawn in the process starting with $g(\tau)$ genomes of each type τ. We then have:*

1. $\mathbb{E}[D^\sigma] = \frac{g(\sigma)}{\langle g, \mathbf{1}\rangle}$.

2. *The following concentration inequality holds for* $\varepsilon \geq 0$*:*

$$\mathbb{P}\left[|D^\sigma - \mathbb{E}[D^\sigma]| > \varepsilon\right] \leq 2\exp\left(-\varepsilon^2 N\right).$$

Similarly for the multivariate binomial distribution, we have the following:

Fact B.5. *Consider N genomes with* $f(\sigma)$ *genomes of each type* σ*. Let* D^σ *be the random variable denoting the fraction of genomes of type* σ *after a mutation step. Then*

1. $\mathbb{E}[D^\sigma] = \frac{1}{N}\sum_\sigma f(\tau) Q_{\tau\sigma}$.

2. *The following concentration inequality holds for* $\varepsilon \geq 0$*:*

$$\mathbb{P}\left[|D^\sigma - \mathbb{E}[D^\sigma]| > \varepsilon\right] \leq 2\exp\left(-2\varepsilon^2 N\right)$$

Fact B.4 is a consequence of Azuma's inequality, and a proof can be found in the book by Dubhashi and Panconesi [DP09]. Fact B.5 is essentially a restatement of the Chernoff-Hoeffding bound. Combining the above bounds, we can deduce the following concentration inequality for each step of the RSM process:

Lemma B.6. *Consider a state* $\mathbf{N_t}$ *of the RSM process. We then have*

1. $\mathbb{E}\left[D^\sigma_{t+1} | \mathbf{N_t}\right] = \frac{(\mathbf{D_t} A Q)_\sigma}{\langle \mathbf{D_t}, A\rangle} = r^\sigma(\mathbf{D_t})$*, with* r^σ *as defined in equation (2).*

2. *Let* ε_1 *and* ε_2 *be arbitrary positive constants. Then with probability (conditional on* $\mathbf{N_t}$*) at least* $1 - 2^{2L+1}\left(\exp\left(-\varepsilon_1^2 N\right) + \exp\left(-2\varepsilon_2^2 N\right)\right)$*, we have* $|D^\sigma_{t+1} - \mathbb{E}\left[D^\sigma_{t+1} | \mathbf{N_t}\right]| \leq (\varepsilon_1 + \varepsilon_2)$ *for every* σ*. In particular, choosing* $\varepsilon_1 = \varepsilon_2 = \varepsilon/2$*, we get that with probability (conditional on* $\mathbf{N}_t$*) at least* $1 - 2^{2L+2}\exp\left(-\varepsilon^2 N/4\right)$*, we have* $|D^\sigma_{t+1} - \mathbb{E}\left[D^\sigma_{t+1} | \mathbf{N_t}\right]| \leq \varepsilon$ *for every* σ*.*

Proof. For ease of notation, let $g_\sigma = a_\sigma N_t^\sigma$. Let I^σ be the random variable denoting the fraction of genomes of type σ left after the selection step. Thus, we have

$$\mathbb{E}[I^\sigma | \mathbf{N}_t] = \frac{g_\sigma}{\langle \mathbf{g}, \mathbf{1}\rangle} \text{ and } \mathbb{E}\left[D^\sigma_{t+1} | \mathbf{I}, \mathbf{N}_t\right] = \sum_\tau I^\tau Q_{\tau,\sigma}.$$

Using a union bound over all genome types with the concentration inequality in Fact B.4, we get that with probability at least $1 - 2^{L+1}\exp\left(-\varepsilon_1^2 N\right)$ conditioned on $\mathbf{N}_t$, we have

$$\left| I^\sigma \in \frac{g_\sigma}{\langle \mathbf{g}, \mathbf{1}\rangle} \right| \leq \varepsilon_1, \text{ for all } \sigma.$$

We denote the above event by $\mathscr{E}$. Now, we consider the concentration of $\mathbf{D}_t$ conditioned on $\mathbf{I}$. Using a union bound over all genome types along with the concentration inequality in Fact B.5, we get with probability at least $1 - 2^{L+1}\exp\left(-2\varepsilon_2^2 N\right)$ conditioned on $\mathbf{I}$, we have

$$\left| D^\sigma_{t+1} - \sum_\tau I^\tau Q_{\tau\sigma} \right| \leq \varepsilon_2, \text{ for all } \sigma.$$

We denote the above event by $\mathscr{F}$. With probability at least $1-2^{2L+1}(\exp\left(-\varepsilon_1^2 N\right)+\exp\left(-2\varepsilon_2^2 N\right))$, conditioned on $\mathbf{N}_t$, both $\mathscr{E}$ and $\mathscr{F}$ occur, and then we have, for all σ,

$$\begin{aligned}\left|D_{t+1}^{\sigma}-\mathbb{E}\left[D_{t+1}^{\sigma}|\mathbf{N}_t\right]\right| &= \left|\sum_{\tau} Q_{\tau\sigma}(D_{t+1}^{\sigma}-\frac{g_\tau}{\langle g,\mathbf{1}\rangle})\right| \\ &\leq \left|D_{t+1}^{\sigma}-\sum_{\tau} I^{\tau}Q_{\tau\sigma}\right|+\sum_{\tau} Q_{\tau\sigma}\left|I^{\tau}-\frac{g_\tau}{\langle g,\mathbf{1}\rangle}\right| \\ &= \varepsilon_2+\varepsilon_1\sum_{\tau} Q_{\tau\sigma}=\varepsilon_1+\varepsilon_2,\end{aligned}$$

which is what we sought to prove. □

Before proceeding, we need the following lemma:

Lemma B.7. *Fix a fitness landscape A with positive entries and a mutation transition matrix Q with $\mu<1/2$. The functions r^{τ} defined in equation (2) are Lipschitz with Lipschitz constant*

$$K=\frac{\max_\tau a_\tau}{\min_\tau a_\tau}\left((1-\mu)^L-\mu^L\right)$$

on the set of probability distributions over genomes.

Proof. For a probability distribution $\mathbf{x}$ over genomes, we have

$$\begin{aligned}\left|\frac{\partial r^{\sigma'}(\mathbf{x})}{\partial x_\sigma}\right| &= \frac{a_\sigma}{\sum_\tau a_\tau x_\tau}\left|Q_{\sigma\sigma'}-r^{\sigma'}(\mathbf{x})\right| \\ &\leq \frac{\max_\tau a_\tau}{\min_\tau a_\tau}\left((1-\mu)^L-\mu^L\right),\end{aligned}$$

where the last line follows by noticing that fact that for all $\mathbf{x}$ and all σ', $\min_{\sigma,\tau} Q_{\sigma\tau}\leq r^{\sigma'}(\mathbf{x})\leq \max_{\sigma,\tau} Q_{\sigma\tau}$, and $\min_{\sigma\tau} Q_{\sigma\tau}=\mu^L$, while $\max_{\sigma\tau} Q_{\sigma\tau}=(1-\mu)^L$. Thus, by the mean value theorem, for any probability distributions $\mathbf{x}$ and $\mathbf{y}$ over genomes, we get

$$|r(\mathbf{x})-r(\mathbf{y})|\leq K\,||\mathbf{x}-\mathbf{y}||_1\,.$$

□

Proof (of Theorem 5.6). Fix a time t_0. In the rest of the proof, we drop the conditioning on the initial state being concentrated on the master sequence for ease of notation. We will prove the following claim by induction for $0\leq t\leq t_0$:

Claim B.8. *For every $\sigma\in\{0,1\}^L$ and $0\leq t\leq t_0$, there exist $l_t^{\sigma},u_t^{\sigma}$ and p_t satisfying the conditions*

1. *$0\leq l_t^{\sigma}\leq u_t^{\sigma}\leq 1$, and $s_t\stackrel{\text{def}}{=}\max_\sigma u_t^{\sigma}-l_t^{\sigma}$ and p_t are $o_N(1)$. Also, m_t^{σ} lies in the interval $[l_t^{\sigma},u_t^{\sigma}]$.*
2. *With probability at least $1-p_t$, D_t^{σ} lies in the interval $[l_t^{\sigma},u_t^{\sigma}]$ for all σ.*

We first see how to finish the proof of Theorem 5.6 assuming Claim B.8. From item 2 in Claim B.8, and using $m_t^{\sigma}\in[l_t^{\sigma},u_t^{\sigma}]$,we get

$$|\mathbb{E}\left[D_{t_0}^{\sigma}\right]-m_{t_0}^{\sigma}|\leq p_{t_0}+(1-p_{t_0})|u_{t_0}^{\sigma}-l_{t_0}^{\sigma}|,\text{ for all }\sigma. \tag{12}$$

Now item 1 of the claim implies that the right hand side of equation (12) goes to 0 as $N\to\infty$, which concludes the proof of Theorem 5.6, assuming Claim B.8. □

We now proceed to prove Claim B.8.

Proof of Claim B.8. At $t = 0$, we can set $l_t^\sigma = u_t^\sigma = m_t^\sigma$, and $p_t = 0$. By the definition of the starting state, this satisfies the conditions claimed in the claim. Now suppose that we have shown that with probability $1 - p_t$, we have $D_t^\sigma \in [l_t^\sigma, u_t^\sigma]$ for all σ. We call the latter event $\mathscr{E}_t$. Recall that

$$\mathbb{E}\left[D_{t+1}^\sigma | \mathbf{D}_t\right] = r^\sigma(\mathbf{D}_t),$$

and define

$$l'^\sigma_{t+1} = \min_{\{\mathbf{y} | y^\sigma \in [l_t^\tau, u_t^\tau]\}} r^\sigma(\mathbf{y}); \;\; u'^\sigma_{t+1} = \max_{\{\mathbf{y} | y^\tau \in [l_t^\tau, u_t^\tau]\}} r^\sigma(\mathbf{y})$$

Notice that $m_{t+1}^\sigma \in \left[l'^\sigma_{t+1}, u'^\sigma_{t+1}\right]$. Also, because of the Lipschitz condition on the function r^σ shown in Lemma B.7, we have $u'^\sigma_{t+1} - l'^\sigma_{t+1} \leq 2^L K s_t$. Now, we condition on the event $\mathscr{E}_t$ defined above, and in this case, we have

$$\mathbb{E}\left[D_{t+1}^\sigma | \mathscr{E}_t\right] \in \left[l'^\sigma_{t+1}, u'^\sigma_{t+1}\right], \text{ for all } \sigma.$$

Choose $\varepsilon(N) = N^{-1/3} = o_N(1)$, and set $l_{t+1}^\sigma = l'^\sigma_{t+1} - \varepsilon/2$, $u_{t+1}^\sigma = u'^\sigma_{t+1} + \varepsilon/2$. Using the concentration result quoted in Lemma B.6, we get that conditioned on $\mathscr{E}_t$, with probability at least $1 - p(N)$ where $p(N) = \exp(-\Omega(N^{1/3})) = o_N(1)$,

$$D_{t+1}^\sigma \in \left[l_{t+1}^\sigma, u_{t+1}^\sigma\right], \text{ for all } \sigma.$$

Now, we saw above that $\mathscr{E}_t$ occurs with probability at least $1 - p_t$. Hence, by a union bound, we get that with probability at least $1 - p_{t+1}$, where $p_{t+1} = p_t + p(N)$,

$$D_{t+1}^\sigma \in \left[l_{t+1}^\sigma, u_{t+1}^\sigma\right], \text{ for all } \sigma.$$

This proves the induction hypothesis, except that we need to make sure that s_t, p_t are $o_N(1)$. We first consider s_t. From above, we have the following recurrence for s_t:

$$s_{t+1} \leq 2^L K s_t + \varepsilon(N); \;\; s_0 = 0. \tag{13}$$

This satisfies $s_t = O_N(\varepsilon) = o_N(1)$ for all $t \leq t_0$, by the choice of ε. Similarly, we have $p_t = t p(N) = o_N(1)$ for $t \leq t_o$ by the choice of $p(N)$. This proves Claim B.8. □

B.3 Proof of Corollary 5.7

We begin by noticing that for $0 < \mu < 1/2$, we can choose a time t_0 such that for $t > t_0$, the state $\mathbf{m}_t$ of the quasispecies model satisfies

$$\left|d^h(\mathbf{m}_t, \mathbf{U}) - d^h(\mathbf{v}_\mu, \mathbf{U})\right| \leq \delta/2, \tag{14}$$

where $\mathbf{v}$ is the unique stationary vector of the quasispecies model. Now fix $t > t_0$. Since the distance function d^h is continuous, Theorem 5.6 allows us to choose an N_δ such that for $N > N_\delta$,

$$\left|d^h(\mathbf{m}_t, \mathbf{U}) - d^h(\mathbb{E}\left[\mathbf{D}_{t,\mu,N}\right], \mathbf{U})\right| \leq \delta/2. \tag{15}$$

Combining equations (14) and (15), we get

$$\left|d^h(\mathbf{v}, \mathbf{U}) - d^h(\mathbb{E}\left[\mathbf{D}_{t,\mu,N}\right], \mathbf{U})\right| \leq \delta.$$

Thus, when $\mu < \mu_c^h(\varepsilon)$, we have

$$d^h\left(\mathbb{E}\left[\mathbf{D}_{t,\mu,N}\right], \mathbf{U}\right) \geq \varepsilon - \delta, \text{ when } \mu < \mu_c^h(\varepsilon), \text{ and,}$$

and when $\mu = \mu_c^h(\varepsilon)$,

$$d^h\left(\mathbb{E}\left[\mathbf{D}_{t,\mu,N}\right], \mathbf{U}\right) \leq \varepsilon + \delta \text{ when } \mu = \mu_c^h(\varepsilon).$$

B.4 Proof of Theorem 5.8

Before proving Theorem 5.8, we first set up some notation for the coupling argument.

Definition B.9. *A coupling of two probability distributions $\mathscr{D}_1$ and $\mathscr{D}_2$ is a pair of random variables (X,Y) defined on a single probability space such that the marginal distribution of X is $\mathscr{D}_1$ and the marginal distribution of Y is $\mathscr{D}_2$.*

Definition B.10. *A coupling of Markov chains with transition matrix $\mathscr{M}$ is defined to be a process $(X_t,Y_t)_{t=0}^{\infty}$ with the property that both (X_t) and (Y_t) are Markov chains with transition matrix $\mathscr{M}$, although the two chains may possibly have different starting distributions.*

Any coupling of Markov chains with transition matrix $\mathscr{M}$ can be modified so that the two chains stay together at all times after their first simultaneous visit to a single state: more precisely, such that if $X_s = Y_s$, then $X_t = Y_t$ for $t \geq s$. In the following, we only consider such couplings. The following well known facts are the basis of coupling based methods for proving mixing time bounds.

Theorem B.11. *Let $\{(X_t,Y_t)\}$ be a coupling satisfying the definition above for which $X_0 = \alpha$ and $Y_0 = \beta$. Let τ_{couple} be the first time the chains meet: $\tau_{\text{couple}} \stackrel{\text{def}}{=} \min\{t : X_t = Y_t\}$. Then*

$$\|\mathscr{M}^t(\alpha,\cdot) - \mathscr{M}^t(\beta,\cdot)\|_{TV} \leq \mathbb{P}\left[\tau_{\text{couple}} > t | X_0 = \alpha, Y_0 = \beta\right].$$

Lemma B.12 (Coupling Lemma)**.** *Let X,Y be random variables defined on a finite sample space Ω and let $\mathscr{C}$ be any coupling of C and Y. Then*

$$\min_{\mathscr{C}} \mathbb{P}_{\mathscr{C}}[X \neq Y] = \|X - Y\|_{TV}.$$

Definition B.13. *Let $d : \Omega \times \Omega \longrightarrow \mathbb{R}_{\geq 0}$ be a distance metric on the state space Ω of the two Markov chains $\{X_t\}_t$ and $\{Y_t\}_t$. Suppose $\mathscr{C}$ is a coupling such that for every $t \geq 0$,*

$$\mathbb{E}[d(X_{t+1},Y_{t+1})] \leq \theta \cdot \mathbb{E}[d(X_t,Y_t)]$$

for every starting distributions X_0, Y_0, then we call $\mathscr{C}$ a (θ,d) coupling. Note that this implies that

$$\mathbb{E}[d(X_t,Y_t)] \leq \theta^t \cdot D,$$

where $D = \max_{\sigma,\tau\in\Omega} d(\sigma,\tau)$.

Fix a integer valued distance function d. Let let $\{X_t\}_t$ be a realization of the Markov chain starting from X_0 and Y_t be another realization starting from the stationary distribution π of the Markov chain. If $\mathscr{C}$ is a (θ,d) coupling, then it follows from the Coupling Lemma that

$$\|X_t - \pi\|_{TV} \stackrel{\text{Coupling}}{\leq} \mathbb{P}[X_t \neq Y_t] \stackrel{d \text{ integral}}{=} \mathbb{P}[d(X_t,Y_t)] \geq 1] \stackrel{\text{Markov}}{\leq} \mathbb{E}[d(X_t,Y_t)] \leq \theta^t \cdot D.$$

This implies that the mixing time $\tau_{\text{mix}}(\varepsilon) = O\left(\frac{\log\ (D/\varepsilon)}{\log\ (1/\theta)}\right)$ when $\theta < 1$.

B.4.1 A Coupling for the RSM Process

The coupling $\mathscr{C}$ we will construct will have two independent parts $\mathscr{C} = (\mathscr{C}_S, \mathscr{C}_M)$, $\mathscr{C}_S$ for the Reproduce-Select phase and $\mathscr{C}_M$ for the mutation part. We first describe the somewhat simpler mutation coupling.

Mutation Coupling $\mathscr{C}_M$. Let $X_t = \{\sigma_1, \ldots, \sigma_N\}$ and $Y_t = \{\nu_1, \ldots, \nu_N\}$. Let M denote an arbitrary permutation on $[N]$ such that $M(i)$ denotes the image of $i \in [N]$. We define the distance between the states as

$$d_{\text{match}}(X_t, Y_t) \stackrel{\text{def}}{=} \min_M \left\{ \sum_{i=1}^{N} d_H\left(\sigma_i, \nu_{M(i)}\right) \right\}.$$

The mutation coupling follows the following algorithm, with M set to be the permutation which achieves the minimum in the above definition.

1. For $i = 1, \ldots, N$
   (a) For $j = 1, \ldots, L$
      i. Choose independently and uniformly at random r_j from $[0,1]$.
      ii. If $\sigma_i(j) = \nu_{M(i)}(j)$
         A. Flip $\sigma_i(j)$ and $\nu_{M(i)}(j)$ if and only if $r_j \geq 1-\mu$.
      iii. Else
         A. Let $r_{j,1} \stackrel{\text{def}}{=} r_j$ and $r_{j,2} \stackrel{\text{def}}{=} 1 - r_j$.
         B. Flip $\sigma_i(j)$ if and only if $r_{j,1} \geq 1-\mu$.
         C. Flip $\nu_{M(i)}(j)$ if and only if $r_{j,2} \geq 1-\mu$.

Lemma B.14. *For $\mu \leq 1/2$, $\mathscr{C}_M$ is a $((1-2\mu), d_{\text{match}})$ coupling.*

Proof. To prove that $\mathscr{C}_M$ is a valid coupling one just needs to note that if r is distributed uniformly at random in $[0,1]$ then so is $1-r$. Hence, for $i = 1, \ldots, N$ and $j = 1, \ldots, L$, each bit $\sigma_i(j)$ (respectively, $\nu_i(j)$)) flips with probability exactly μ. Further, these flips are independent by construction.

To prove that $\mathscr{C}_M$ is a $((1-2\mu), d_{\text{match}})$ coupling, let X_t, Y_t be the states of the two Markov chains with distance $d \stackrel{\text{def}}{=} d_{\text{match}}(X_t, Y_t)$. By definition of d_{match}, there is some matching $M^\star$ which achieves d. Without loss of generality assume that $M^\star$ is identity, i.e., $M^\star(i) = i$, for all $1 \leq i \leq N$. Hence, $\sum_{i=1}^{N} d_H(\sigma_i, \nu_i) = d$. Let $X_{t+1} \stackrel{\text{def}}{=} (\sigma_1^{t+1}, \ldots, \sigma_N^{t+1})$ and $Y_{t+1} \stackrel{\text{def}}{=} (\nu_1^{t+1}, \ldots, \nu_N^{t+1})$ be the output of $\mathscr{C}_M$ on input $(\sigma_1, \ldots, \sigma_N)$ and $(\nu_1, \ldots, \nu_N)$ respectively. We will calculate $\mathbb{E}\left[\sum_{i=1}^{N} d_H(\sigma_i^{t+1}, \nu_i^{t+1})\right]$ and show that it is exactly $(1-2\mu) \cdot d$. Hence, $d_{\text{match}}(X_{t+1}, Y_{t+1}) \leq (1-2\mu) \cdot d$, as d_{match} is defined as the minimum over all possible matchings. By linearity of expectation it is sufficient to show that for all $i = 1, \ldots, N$,

$$\mathbb{E}\left[d_H(\sigma_i^{t+1}, \nu_i(t+1)\right] = (1-2\mu) \cdot d_H(\sigma_i, \nu_i).$$

Again by linearity of expectation it is sufficient to show the following:

$$\mathbb{E}_{r_j}\left[d_H(\sigma_i^{t+1}(j), \nu_i^{t+1}(j))\right] = (1-2\mu) \cdot d_H(\sigma_i(j), \nu_i(j)).$$

This follows from observing that if $\sigma_i(j) = \nu_i(j)$, then $\mathbb{P}\left[\sigma_i^{t+1}(j) = \nu_i^{t+1}(j)\right] = 1$, while if $\sigma_i(j) \neq \nu_i(j)$, then, as $\mu \leq 1/2$, $\mathbb{P}\left[\sigma_i^{t+1}(j) = \nu_i^{t+1}(j)\right] = 2\mu$. Hence, $\mathbb{P}\left[\sigma_i^{t+1}(j) \neq \nu_i^{t+1}(j)\right] = 1-2\mu$. This completes the proof.

□

Coupling $\mathscr{C}_S$ for the Selection Process. We again consider two states $X_t = \{\sigma_1, \sigma_2, \ldots, \sigma_N\}$ and $Y_t = \{\nu_1, \nu_2, \ldots, \nu_N\}$. Our distance function is still d_{match} defined above. We first note the $d_{\text{match}}(\cdot,\cdot)$ is actually a metric.

Lemma B.15. *$d_{\text{match}}(\cdot,\cdot)$ is a metric.*

Proof. By construction $d_{\text{match}}(X,Y) \geq 0$ with equality if and only if $X = Y$. Now consider states $X = \{\sigma_i\}_{i=1}^N, Y = \{\tau_i\}_{i=1}^N$ and $Z = \{\nu_i\}_{i=1}^N$ in the state space Ω. Let α and β be permutations of $[N]$ such that

$$d_{\text{match}}(X,Y) = \sum_{i=1}^N d_H\left(\sigma_i, \tau_{\alpha(i)}\right) \text{ and } d_{\text{match}}(Y,Z) = \sum_{i=1}^N d_H\left(\tau_i, \nu_{\beta(i)}\right)$$

Now we have

$$\begin{aligned} d_{\text{match}}(X,Z) &\leq \sum_{i=1}^N d_H\left(\sigma_i, \nu_{\beta(\alpha(i))}\right) \\ &\leq \sum_{i=1}^N \left(d_H\left(\sigma_i, \tau_{\alpha(i)}\right) + d_H\left(\tau_{\alpha_i} + \nu_{\beta(\alpha(i))}\right)\right) \\ &= d_{\text{match}}(X,Y) + d_{\text{match}}(Y,Z). \end{aligned}$$

□

We will use the following general path coupling result of Bubley and Dyer [BD97] to define the coupling $\mathscr{C}_S$.

Theorem B.16 (Path Coupling [BD97])**.** *Consider a Markov chain M on state space Ω and a distance function d on Ω such that $d'(x,x) = 0$ for all $x \in \Omega$. Consider a connected undirected graph G on Ω such that the length of each edge $\{x,y\}$, if present in G, is $d(x,y)$, and let d' be the shortest path metric on G. Suppose there exists a coupling $\mathscr{C}$ for M such that for some $\alpha < 1$, and all $X_t, Y_t \in \Omega$ which are adjacent in G,*

$$\mathbb{E}_{\mathscr{C}}\left[d(X_{t+1}, Y_{t+1}) | X_t, Y_t\right] \leq \alpha d(X_t, Y_t).$$

If every edge of G is a shortest path under the metric d' described above, then the coupling $\mathscr{C}$ can be extended to a coupling $\mathscr{C}'$ such that

$$\mathbb{E}_{\mathscr{C}'}\left[d'(X_{t+1}, Y_{t+1} | X_t, Y_t\right] \leq \alpha d'(X_t, Y_t).$$

for all $X_t, Y_t \in \Omega$.

We will first show now that the path metric resulting from an application of the above theorem to d_{match} is d_{match} itself, since this is crucial for composing the $\mathscr{C}_S$ coupling with the coupling $\mathscr{C}_M$ described above.

Lemma B.17. *Consider the state space Ω of the RSM Markov chain $\mathscr{M}$. Let G be the graph on Ω in which two vertices X and Y are adjacent if and only if $d_{\text{match}}(X,Y) = 1$. Then the path metric d' constructed in Theorem B.16 is identical with d, and each edge in G is a shortest path.*

Proof. For brevity we will denote $d_{\text{match}}(\cdot,\cdot)$ by d. Notice that since each edge is of length 1, it is also a shortest path by construction. Since d is a metric, we also have $d'(X,Y) \geq d(X,Y)$ for all $X, Y \in \Omega$. We now proceed by induction to show that $d(X,Y) \geq d'(X,Y)$ for all $X, Y \in \Omega$. Notice that when $d(X,Y) = 1$,

this is true by definition of d'. Now, suppose that $d(X,Y) \leq k-1$ implies $d'(X,Y) \leq d(X,Y)$, and consider the case $d(X,Y) = k > 1$. We claim that there exists a Z such that $d(X,Z) \leq k-1$ and $d(Y,Z) \leq 1$. The existence of such a Z implies using the induction hypothesis that

$$\begin{aligned} d'(X,Y) &\leq d'(X,Z) + d'(Z,Y) \\ &\leq d(X,Z) + d(Z,Y) \leq k = d(X,Y). \end{aligned}$$

It only remains to construct such a Z. Let $X = \{\sigma_i\}_{i=1}^N$ and $Y = \{\nu_i\}_{i=1}^N$. Without loss of generality, we may assume that $d(X,Y) = \sum_{i=1}^N d_H(\sigma_i, \nu_i)$. Since $d(X,Y) = k > 1$, there exists a j such that $d_H(\sigma_j, \nu_j) \geq 1$. Let s be a string obtained by flipping a single bit of ν_j such that $d_H(\sigma_j, s) = d_H(\sigma_j, \nu_j) - 1$. Now let $Z = \{\tau_i\}_{i=1}^N$, where $\tau_i = \nu_i$ for $i \neq j$ and $\tau_j = s$. By construction, $d(Y,Z) \leq 1$ and $d(X,Z) \leq k-1$.

□

Claims about the Coupling. Suppose we find a coupling C for the selection phase such that when $d_{\text{match}}(X_t, Y_t) = 1$, then the intermediate states $I(X_t)$ and $I(Y_t)$ satisfy

$$\mathbb{E}\left[d_{\text{match}}(I(X_t), I(Y_t)) \,|X_t, Y_t\right] \leq \alpha,$$

then using Theorem B.16 and the coupling for the mutation phase described above, we get

$$\mathbb{E}\left[d_{\text{match}}(X_{t+1}, Y_{t+1} | X_t, Y_t)\right] \leq \alpha(1-2\mu) d_{\text{match}}(X_t, Y_t).$$

This will give us fast mixing as long as $\alpha(1-2\mu) < 1$.

We now describe such a coupling for the selection process, for general X_t and Y_t, which we will analyze only in the simple but sufficient case when $d_{\text{match}}(X_t, Y_t) = 1$. Suppose that X_t and Y_t contain, respectively, n_σ^x and n_σ^y genomes of type σ. After reproduction, the number of genomes of type σ in the two chains is $a_\sigma n_\sigma^x$ and $a_\sigma n_\sigma^y$ respectively. Let the total number of genomes be $M_x = \sum_\sigma a_\sigma n_\sigma^x$ and $M_y = \sum_\sigma a_\sigma n_\sigma^y$ respectively. Without loss of generality let us assume that $M_x \geq M_y$, and set $M = M_x$.

We now construct a bag of M balls as follows. For each σ, the bag has exactly $a_\sigma \min\left(n_\sigma^x, n_\sigma^y\right)$ balls with label $(\sigma, (x,y))$. If $n_\sigma^x \geq n_\sigma^y$, then the bag has exactly $a_\sigma(n_\sigma^x - n_\sigma^y)$ balls with label (σ, x), otherwise it has exactly $a_\sigma(n_\sigma^y - n_\sigma^x)$ balls with label (σ, y). Thus the total number of balls in the bag is M.

We now take a random permutation of the M balls, and take the intermediate state I_X (respectively, I_Y) to be the multiset of genomes given by the first N balls carrying the label (x,y) or x (respectively, (x,y) or y). Notice that a ball carrying a label (x,y) can contribute a genome to both I_X and I_Y.

Claim B.18. *The above coupling is a valid Markovian coupling for the selection phase of the RSM chain.*

Proof. Notice that sampling without replacement a objects from a set of b objects is equivalent to taking the first a elements from a uniform random permutation of the b objects. Also note that given a subset S of a set of b objects, and a uniform random permutation α over the b objects, the restriction of α to the elements of S is a uniformly random permutation of the elements of S. Now consider the set of M labeled balls constructed above, and define S_X (respectively, S_Y) to be the set of balls carrying a (x,y) or x (respectively, $x(x,y)$ or y) label. By the observations above, see that the set I_X (respectively, I_Y) has the same distribution as if it was sampled without replacement from S_X (respectively, S_Y). This proves the claim. □

Lemma B.19. *Suppose $d_{\text{match}}(X_t, Y_t) = 1$. Then under the above coupling, we have*

$$\mathbb{E}\left[d_{\text{match}}(I_X, I_Y) \,|X_t, Y_t\right] \leq \frac{1}{1 - \frac{1}{N}} \frac{\max a_\sigma}{\min a_\sigma} L.$$

Proof. For brevity, we will denote X_t and Y_t as $X = \{\sigma_i\}_{i=1}^N$ and $Y = \{\nu_i\}_{i=1}^N$ and $d_{\text{match}}(\cdot,\cdot)$ by d. Since $d(X,Y) = 1$, we can assume without loss of generality that $\sigma_i = \nu_i$ for $i > 1$, and σ_1 and ν_1 differ in exactly one bit.

We now consider the coupling described above, and let S be the set of balls with label (x,y). Notice that $|S| = \sum_{1<i\leq N} a_{\sigma_i}$. Notice that $|S| \geq (N-1)\min_\sigma a_\sigma$. We assume without loss of generality that $a_{\sigma_1} > a_{\nu_1}$, so that the number of balls M in the bag is $|S| + a_{\sigma_1}$. Consider a random permutation α of the M balls, and let I be the (random) multiset of balls with label (x,y) occurring in the first N positions of α. Notice that $|I_X \cap I_Y| \geq |I|$. We observe that if the intersection of I_X and I_Y (seen as multisets) is of size at least $|I|$, then $d_{\text{match}}(I_X, I_Y) \leq L(N - |I|)$, hence we have

$$\mathbb{E}\left[d_{\text{match}}(I_X, I_Y) \,|X,Y\right] \leq L(N - \mathbb{E}\left[|I| \;|X,Y\right]). \tag{16}$$

Now, we have

$$\begin{aligned}
\mathbb{E}\left[|I| \;|X,Y\right] &= \frac{|S|}{|S| + a_{\sigma_1}} N \\
&\geq N\left(1 - \frac{a_{\sigma_1}}{|S|}\right) \geq N\left(1 - \frac{a_{\sigma_1}}{(N-1)\min_\tau a_\tau}\right).
\end{aligned}$$

Plugging this into equation (16), we get

$$\begin{aligned}
\mathbb{E}\left[d_{\text{match}}(I_X, I_Y) \,|X,Y\right] &\leq \frac{LNa_{\sigma_1}}{(N-1)\min_\tau a_\tau} \\
&\leq \frac{1}{1-\frac{1}{N}} \frac{\max_\tau a_\tau}{\min_\tau a_\tau} L,
\end{aligned}$$

which establishes our claim. □

Using the above discussion, we get fast mixing under the condition that

$$(1-2\mu)\frac{1}{1-\frac{1}{N}} \frac{\max_\sigma a_\sigma}{\min_\sigma a_\sigma} L < 1.$$

Formally, we have

Theorem B.20. *Fix a mutation rate $\mu < 1/2$, and a fitness landscape A. Define*

$$K(A,\mu) = (1-2\mu)\frac{1}{1-\frac{1}{N}} \frac{\max_\sigma a_\sigma}{\min_\tau a_\tau} L.$$

When $K(A,\mu) < 1$, we have $\tau_{\text{mix}}(\varepsilon) = O\left(\frac{\log(NL/\varepsilon)}{\log(1/K)}\right)$.

We note here that the main difficulty in designing coupling arguments for the RSM process is the distance expansion property of the reproduction and selection phases. Given two states of the RSM process, the reproduction phase amplifies the distance between them, and the nature of the selection phase tends to keep this distance intact. In this setting, the chain can be described as a noisy random walk on a boolean hypercube, and our bound reflects the intuition that when the noise is small, the fast mixing property of the hypercube should able to enforce fast mixing of the RSM chain. We reemphasize that we consider that achieving a better understanding of the mixing properties of the RSM walk, in terms of both upper and lower bounds appears to be an interesting and challenging open problem.

B.5 Proof of Theorem 5.9

We now proceed to prove Theorem 5.9. We will use notation similar to that used in the proof of Lemma 5.2 in Appendix B.1. By a slight abuse of notation, we will use P and R for the transition matrices of the projected versions of the chains described in the proof of Lemma 5.2 above. We note that for any numbers M and $M'_1, M'_2, \ldots, M'_l$ summing up to M, $\binom{M}{M'_1, M'_2, \ldots, M'_l}$ is of size at most $O(M \log M)$ bits and can be computed in at most $O(M)$ operations over numbers of size at most $O(M \log M)$. We start with an estimation of the time required to compute the entries of P. Using the above observation, and the form of entries of P, we have the following observation for P.

Observation B.21. *Each entry of P is of size at most $\tilde{O}(N \max a_\sigma)$, and can be computed in $N \max a_\sigma$ operations over integers of size $\tilde{O}(N \max a_\sigma)$.*

We now proceed to estimate the complexity of computing entries of the matrix R. We first get a bound on the time required to pre-compute the $(L+1) \times (L+1)$ matrix $\mathscr{Q}$ defined in the description of the mutation chain in Appendix B.1.

Observation B.22. *Let b be the number of bits required to represent μ. We have*

$$\mathscr{Q}_{st} = (1-\mu)^L \left(\frac{\mu}{1-\mu}\right)^{s-t} \sum_x \binom{L-s}{x} \binom{s}{t-x} \left(\frac{\mu}{1-\mu}\right)^{2x}.$$

Hence, each entry of $\mathscr{Q}$ is of size at most $\tilde{O}(bL)$ and can be pre-computed in time $O(L^2)$ operations over numbers of size $\tilde{O}(bL)$.

Proof of Theorem 5.9. Notice that the number of terms in the sum in equation (8) for the computation of the entry of the matrix $R([h],[h'])$ is at most

$$\begin{aligned}
\prod_{i=0}^{L} \binom{h(i)+L}{L} &\leq \left(\frac{e}{L}\right)^{L(L+1)} \prod_{i=0}^{L} (h(i)+L)^L \\
&\leq \left(e\left(1+\frac{N}{L(L+1)}\right)\right)^{L(L+1)}, \text{ by the AM-GM inequality.}
\end{aligned}$$

We will use the shorthands $S = \binom{N+L}{L} \leq \frac{(N+L)^L}{L!}$ and

$$G = \left(e\left(1+\frac{N}{L(L+1)}\right)\right)^{L(L+1)}.$$

Notice that R is of dimension at most $S \times S$. Computing the products of all the $\mathscr{Q}_{ij}$'s in each of these terms takes N multiplications on numbers of size at most $\tilde{O}(bL)$, and hence produces a number of size $\tilde{O}(NbL)$ in time at most $\tilde{O}(NbL)$. Computing the products of all the multinomial coefficients in each of the terms takes $O(N)$ multiplications on integers of size $\log N$, thus producing an integer of size $O(N \log N)$ in time $\tilde{O}(N \log N)$. The total size of each entry of R is thus at most $s_1 = \tilde{O}(\log G + N \log N + NbL)$, and the entry can be computed in time $t_1 = \tilde{O}(Gs_1)$. All the entries of R can thus be computed in time $S^2 t_1$.

Notice that $\mathscr{M}_w = PR$, and given the above estimates on the sizes of the entries of P and R and the times required to compute their entries, each entry of $\mathscr{M}_w$ is of size at most $s_2 = O(\log S + s_1 + N \max a_\sigma)$, and

can be computed in time $\tilde{O}(Ss_2)$, given the entries of P and R. Thus, the time taken for the computation of $\mathscr{M}_w$, including the computation of R and P and the pre-computation of $\mathscr{Q}$ is

$$T_1 \quad = \quad \tilde{O}\left(bL^3 + S^3 s_2 + S^2 G s_1\right)$$

and each entry of $\mathscr{M}_w$ is of size

$$s_2 = \tilde{O}(NbL + N\max a_\sigma + L^2).$$

The time required to compute an exact solution of $\mathscr{M}_w \pi_w = \pi_w$ with the restriction $||\pi_w||_1 = 1$ using Gaussian elimination is thus of the order $\tilde{O}(S^3 s_2)$. Comparing this with T_1, we get that the total running time is $\tilde{O}(T_1)$, which is what we sought to prove. □

B.6 Proof of Theorem 5.10

INPUT: An initial state $\alpha_0 \in \Omega_w$, an $\varepsilon > 0$, grid resolution δ, accuracy parameter δ_1, and error probability δ_2, L, N and an A which is class invariant.

GOAL: To estimate $\mu_c^h(\varepsilon, N)$.

OUTPUT: μ_0 such that$\mu_0 \geq \mu_c^h(\varepsilon + \delta_1/2)$ and $d^h(\mathbf{D}_{\infty,\mu_0-\delta}, \mathbf{U}) \geq \varepsilon - \delta_1/2$.

Let $\varepsilon' = \frac{\delta_1}{2L^2}$ (the distance from stationarity), $c = \left\lceil \frac{1}{2\delta} \right\rceil$ and $s = \left\lceil \frac{8L^4}{\delta_1^2} \log(2c(L+1)/\delta_2) \right\rceil$ (number of samples from the distribution).

For $\delta \leq \mu \leq 1/2$ in steps of δ,

1. Let $\tau \stackrel{\text{def}}{=} \tau(L, N, A, \mu, \varepsilon')$.
2. Let $\mathbf{D}_{\tau,k}$, for $k = 1, \ldots, s$, denote s independent samples from the RSM process with parameters L, N, A and μ starting from the initial state α_0.
3. Let $\mathbf{Z}_s \stackrel{\text{def}}{=} \frac{1}{s} \sum_{k=1}^{s} \mathbf{D}_{\tau,k}$.
4. **If** $d^h(\mathbf{Z}_s, U) \leq \varepsilon$ then **Return** μ and **Stop**.
5. **Else** $\mu = \mu + \delta$.

Figure 1: The Algorithm ERRORTHRESHOLD

In this section, we give a proof of Theorem 5.10. Recall that π_w denotes the stationary distribution of the projected RSM chain described in Section 5.1. Using the definition of the mixing time, and the Chernoff-Hoeffding bound, we have the following lemma in order to bound the number of samples s required in each iteration of the algorithm:

Lemma B.23. *Suppose we take s samples from independent realizations of the fully mixed projected RSM process, denoting the samples so obtained as $\mathbf{D}(i)$ for $i = 1, 2, \ldots, s$. For any fixed genome σ, we then have $\mathbb{E}[D^\sigma(i)] = \mathbb{E}_{\pi_w}[D^\sigma]$ for all i. Now for every $\varepsilon > 0$*

$$\mathbb{P}\left[\left|\frac{1}{s}\sum_{i=1}^{s} D^\sigma(i) - \mathbb{E}_{\pi_w}[D^\sigma]\right| \geq \varepsilon\right] \leq 2\exp\left(-2\varepsilon^2 s\right).$$

In particular, if we run s independent realizations of the chain up to the time $\tau(L,N,A,\mu,\varepsilon/2)$, and let $\mathbf{D}_i$ denote the sample obtained by the ith realization, then

$$\mathbb{P}\left[\left|\frac{1}{s}\sum_{i=1}^{s} D^{\sigma}(i) - \mathbb{E}_{\pi_w}[D^{\sigma}]\right| \geq \varepsilon\right] \leq 2\exp\left(-\varepsilon^2 s/2\right).$$

Proof of Theorem 5.10. Consider the random variables $\mathbf{Z_s}$ in any of the at most c iterations in ERRORTHRESHOLD. By the choice of ε' and s, we get from the above lemma and a union bound that all the c different variables $\mathbf{Z_s}$ that we get across all the iterations of the loop satisfy

$$||\mathbf{Z_s} - \mathbb{E}_{\pi_w}[\mathbf{D}]||_{\infty} \leq \delta_1/(2L^2) \tag{17}$$

with probability at least $1-\delta_2$. In the rest of the proof, we condition on the above event occurring.

Since the average Hamming distance can be at most L, and we are running the projected chains till $\tau(L,N,A,\mu,\delta_1/(2L^2))$ for each μ, we get that for every μ considered by the algorithm

$$\left|d^h(\pi_{w,\mu},\mathbf{U}) - d^h(\mathbf{Z_s},\mathbf{U})\right| \leq \delta_1/2.$$

We therefore deduce that when the algorithm outputs a μ, $d^h(\pi_{w,\mu},\mathbf{U}) \leq \varepsilon + \delta_1/2$, using the conditioning on the event in equation (17). Similarly, by noticing that the algorithm had $d^h(\mathbf{Z_s},\mathbf{U}) \geq \varepsilon$ for $\mu - \delta$, we get that $d^h(\pi_{w,\mu-\delta},\mathbf{U}) \geq \varepsilon - \delta_1/2$. The estimate of the running time follows by noticing that assuming bits with bias μ can be sampled in $O(\log(1/\mu))$ time, it takes time $O(NL\max a_\sigma \log(1/\mu))$ to simulate each step of the $\mathscr{M}_w$ chain. The bound on the error probability follows from the conditioning used on the validity of equation (17). □

We comment briefly on a technical point about the definition of the error threshold that has been used in the literature (and that we use too). With this definition, there might exist a μ satisfying $1/2 > \mu > \mu_c^h(\varepsilon)$ such that $d^h(\mathbf{v},\mathbf{U}) > \varepsilon$. An analogous condition might hold in the finite population case too. If we could preclude the occurrence of such anomalous behavior of error-thresholds, we would be able to improve the guarantee on the output μ_0 of the algorithm ERRORTHRESHOLD to be of the form $\mu_c^h(\varepsilon+\delta_1,N) \leq \mu_0 \leq \mu_c^h(\varepsilon-\delta_1)+\delta$. We observe, however, that somewhat surprisingly, even for the simpler case of the quasispecies model, to the best of our knowledge, no attempts have been made to rigorously prove that such anomalous behavior cannot occur. We leave the resolution of this point for the finite population case as an open problem.